Genetics and Epigenetics of Psychiatric Disorders

by

Chang(April) Shu

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

April, 2018

Abstract

Genetics and epigenetics studies have been used widely for psychiatric and behavior disorders such as substance use and Autism Spectrum Disorders(ASD). By utilizing genetics on ASD and epigenetics data on injection drug use(IDU) from the Department of Mental Health, we are able to identify genetic or epigenetic region of interests. In Aim 1 and Aim 2, we used an ongoing longitudinal data called the AIDS Linked to the Intravenous Experience (ALIVE) study, and it contained past six month IDU phenotypic data and epigenetics data from peripheral blood on 2-4 visits from 288 subjects. Taking advantage of the longitudinal design, we conducted four separate epigenome-wide association analyses(EWAS) on past six month any injection drug use, heroin only injection use, cocaine only injection use, and co-use of heroin and cocaine injection. To borrow information across these four separate EWASs based on correlated phenotypes, we modified the correlation motif method and applied it into epigenetics data, and the top hits after joint analyses by correlation motif are now epigenome-wide significant after adjusting for multiple comparison. The epigenetic marker near the FKBP5 gene was found to be associated with IDU,

which is also a gene of interest for other types of psychiatric disorders. In the ALIVE study, we also conducted EWAS on HIV, and used PCs that are negatively associated with CD4+ cells and positively associated with CD8+ cells to account for cell composition difference between chronic HIV infected individuals and HIV negative individuals. We discovered a epigenome-wide significant methylation site near NLRC5 gene, which is also reported in an independent EWAS study on HIV. In another genetic study called the Study to Explore Early Development (SEED), we have ASD phenotypic information and genetics information on about 1200 cases and controls. We utilized brain expression quantitative loci(brain eQTLs) from public literature to extract brain eQTL single nucleotide polymorphism(brain eSNPs) to reduce the search space from genome-wide variation to only brain expression related SNPs. We discovered that only temporal cortex eSNPs shows qqplots that is of suggestive association, and other brain region eSNPs are significantly deflated. This finding confirms with other imaging studies on temporal cortex might be affected by ASD.

Primary Reader: Brion Maher

Acknowledgments

I would like to thank my thesis advisors Dr. Brion Maher, Dr. Dani Fallin, and Dr. Hongkai Ji. Dr. Brion Maher has been a great mentor throughout my PhD at Johns Hopkins. I enjoyed working as well as talking about life with him. He is very open-minded and has provided supports from all aspects. I cannot survive my PhD without his strong support. Dr. Dani Fallin is a brilliant PI to work with, and I always admire her working efficiency and her thoughtful ideas on my projects. Dr. Hongkai Ji has been extremely helpful regarding with any statistical issues and has provided insightful advice to my projects. I couldn't thank him more for being my MHS advisor in biostatistics and being in my committee chair in numerous occasions.

In my committee, Dr. Andrew Jaffe, Dr. Christine Ladd-Acosta and Dr. Kelly Benke also provided great mentoring. Dr. Andrew Jaffe is extremely smart and I really appreciate his ideas on how to handle the cell composition problems. He also introduced many cutting-edge papers in this field. Dr. Chirstine Ladd-Acosta has mentored me over the very first project in my PhD, and she helped me understand how to conduct genetics and epigenetics projects. Dr. Kelly Benke has been really

supportive when I have any questions or issues during my PhD. She not only provides thoughts and ideas on projects but also on any issues that I have for my PhD and career. Dr. Greg Kirk and Dr. Shruti Mehta have provided valuable insights on the phenotype data and help me understand the data well.

I really appreciate the great efforts from departmental staffs especially Patty Scott and Michelle Maffett. Patty is always there for me whatever question that I have, and I couldn't appreciate it more. Michelle is very responsible, and really help me on scheduling important meetings.

I also want to thank my peers in the School of Public Health. Shan Andrews helped me a lot with his extensive experience in epigenetic data, and Weiyan Li shared valuable experience with me. I learned a lot from Dani Sisto who is really good at coding. I couldn't list everyone who encouraged and generously offered their help, but it is a great honor to get to know brilliant fellow students through the ALIVE study and SEED study, from Department of Mental Health, Department of Epidemiology and Department of Biostatistics. I especially want to thank our "wolf pack" cohort, which is an extremely supportive PhD group. I'm also extremely grateful to Weiling Zhou, Qingyuan He, Tong Qu, and etc., who are always there for me whenever I need.

I couldn't have the opportunity to go to Johns Hopkins without my family's support. Lastly, I want to thank my fianc Jingjing and Haohao, who have always got my back from high school, college, graduate school and PhD.

Dedication

This thesis is dedicated to Jingjing and Haohao, who have been supporting me for many years and more.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

1.1.1 Why mental health and psychiatric genetics?

The prevalence of mental disorders and global burden has often been underestimated. The lifetime prevalence of common mental disorders is 29.2% (25.9-32.6%) according to a meta analysis from 1980-2013.[3] In the US, about half of the sample in a nationally representative survey met lifetime diagnostic criteria for a mental disorder.[4] In addition to the high prevalence of mental disorder, the economic cost and burden is substantial,[5] including disability, premature mortality, burden to caregiver, etc. Recent research has argued that the true burden of mental illness is underestimated, and the global burden of mental disorders ranks first among all physical and

mental illness in terms of years lived with disability.[6]

Despite the high prevalence and tremendous burden to patients, patient's families and the society, the underlying biological mechanism of most mental disorders remains mostly unknown. For example, in Autism Spectrum Disorder (ASD), the current treatment options are limited to relieving symptoms but cannot fully cure the disease.[7] Another challenge is that the diagnosis of mental disorders is more complicated than physical illness. Unlike type 2 diabetes and heart diseases, there rarely exist valid biomarkers as valid measures for mental disorders.

The current challenges in understanding mental health disorders and treatment reflects our lack of knowledge of the brain. The brain plays the most important role in our body since it acts as the central processing unit to regulate our body in response to environmental changes. It is also the most complex tissue to study and most difficult tissue to obtain from humans. It is not surprising that unlike diseases affecting other tissues, there have been fewer advances in disorders related to brain.

However, the current advances in genetic tools open a window to study brain related genes. The genotyping technology has been improved dramatically during the past decade, and the cost is greatly lowered, leading to numerous genetics studies on large populations. Genetics studies are very important to mental health research since it has been long observed that some mental disorders aggregate in families and genes may be involved in many mental disorders. Since the etiology of most mental disorders remains unknown, genetics studies provide a method to pinpoint proteins

and enzymes encoded in the genome that implicate specific biological pathways.

The Department of Mental Health in Johns Hopkins Bloomberg School of Public Health provides great opportunities to study the genetics and epigenetics of mental health disorders, and there are two great sources of genetics and epigenetics data on substance use from Dr. Brion Maher and Autism Spectrum Disorders from Dr. Dani Fallin. I'm more than honored to have the opportunity to learn to analyze these valuable data and contribute to our collective knowledge on the etiology of mental disorders.

The following background section will provide background knowledge on general genetics and epigenetics tool used in psychiatric genetics, epidemiology and genetic/epigenetic advances in injection drug use and autism spectrum disorder.

1.1.2 Genetics and epigenetics approaches in mental health

1.1.2.1 Genome-wide association study (GWAS)

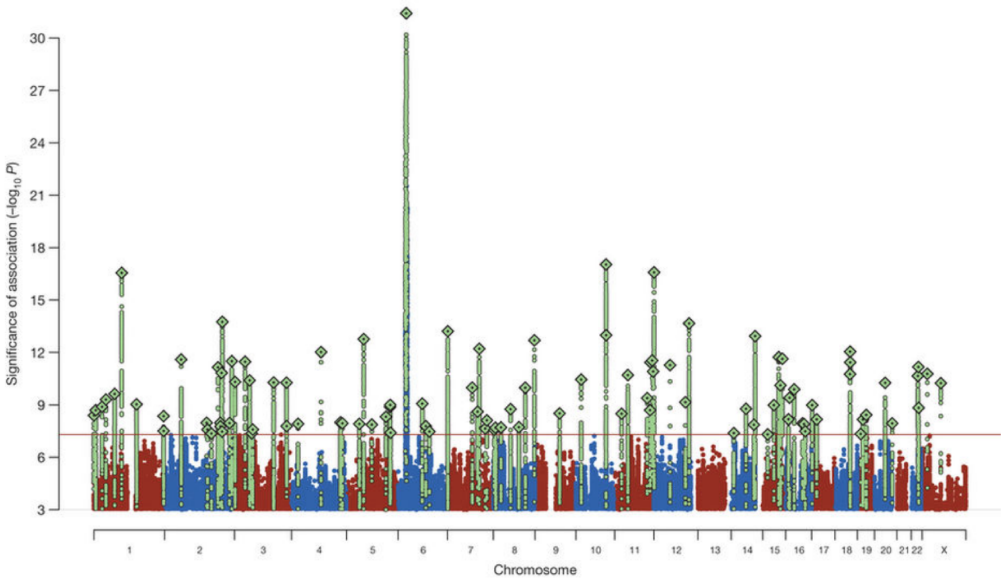
The genome-wide association study(GWAS) focuses on finding the association between the phenotype of interest and genetic variants across the genome.[8] The unit of genetic variation GWAS is called single nucleotide polymorphism or SNP,[8] referring to one unit base-pair variation in a specific genomic location. The number of detectable SNPs ranges from 1 million to 10 million depending on the depth of

genotyping technology used. The genotyping technology has also evolved rapidly and the cost of genotyping an individual has become much more affordable in recent years.

Large consortia have formed to provide power to scan millions of SNPs without any prior hypothesis or knowledge. GWAS using large populations to find disease related genetic variants took off in 2007. For example, the Wellcome Trust consortium from UK conducted a GWAS study on 14,000 cases of seven diseases.[9] With a substantial number of GWAS published, methods on interpreting and prioritizing GWAS results,[10] and reflections on what GWAS can contribute to our knowledge have resulted.[11]

In psychiatric disorders, several genome-wide association studies (GWAS) have been carried out for five major psychiatric disorders,[12] and have been most successful in schizophrenia with 108 genetic loci and more to come.[1]

GWAS study on psychiatric disorders provide many biologically meaningful entry points for further investigation. Many significant genetic variants across psychiatric disorders have been shown to be enriched for SNPs that associate with expression levels, particularly in brain.[12,13] These functional genetic variants and related genes provide insights on the underlying biological mechanism. In schizophrenia, a GWAS informed strong genetic variant is found in the major histocompatibility complex locus.[14] The downstream biological studies showed that this variant in involved in synapse elimination during brain development and sheds some insight on the biological process related to schizophrenia.[14]



Manhattan plot of the discovery genome-wide association meta-analysis of 49 case control samples (34,241 cases and 45,604 controls) and 3 family based association studies (1,235 parent affected-offspring trios). The *x* axis is chromosomal position and the *y* axis is the significance ($-\log_{10}$ *P*; 2-tailed) of association derived by logistic regression. The red line shows the genome-wide significance level (5×10^{-8}). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.

Figure 1.1: Manhattan plot showing significant hits in schizophrenia GWAS[1]

With schizophrenia as an example showing how genetics research like GWAS can help inform its etiology, mental disorders such as Autism Spectrum Disorders (ASD) and Substance Use Disorders (SUD) can possibly benefit from genetics studies with larger sample sizes and careful data harmonization.

1.1.2.2 Epigenome-wide association study (EWAS)

Studies on mental and behavior disorders have shown that most mental health disorders are a combination of both genes and environment.[15–17] The DNA sequence mostly remains unchanged over time, but epigenetic profiles on top of the DNA sequence, such as DNA methylation and histone modification might change in response

to environmental exposures.[18] Thus,Epigenetics is an important area to the study gene and environment interaction.

Epigenetics is defined as all heritable changes in gene expression that are not coded in the DNA sequence itself, and DNA methylation is one of the most commonly measured epigenetic markers.[19] A recent review paper on substance use disorder has summarized several important CpG sites associated with substance use such as alcohol and tobacco use in animal and human studies.[20] It was also shown that DNA methylation patterns from 26 CpG sites at 3-5 years of age can accurately classify prenatal exposure to smoking.[21] There is also literature suggesting DNA methylation may play a role in Autism Spectrum Disorders(ASD) and other developmental disorders.[22,23]

A goal of epigenetics research is the identification of markers that can serve as either biomarkers for environmental exposure, or starting points for downstream etiology research. Thus, just like GWAS, epigenome-wide association studies (EWAS) can also shed light on psychiatric disease. By scanning possible associations between phenotype and 450,000 to 850,000 DNA methylation markers across epigenome, biologically interesting epigenetics markers may come to light.

1.1.3 Injection Drug Use(IDU) and HIV

1.1.3.1 Epidemiology

There are estimated to be 11.0-21.2 million injection drug users (IDUs) worldwide in 2007, with HIV prevalence over 40% among IDUs in nine countries.[24] The prevalence of substance use is high in the US, with prevalence of 13.8% in 2012-2014 in large metropolitan areas in a national survey of reported past-month use of any illicit substance.[25] According to the most recent results on National Survey of Drug Use and Health, there were about 475,000 people aged 12 or older who were injecting heroin in 2016 in the US.[26] Estimates of cocaine use in 2016 were approximately 1.9 million people in the US.

Baltimore has had significant drug use problems for many years, and the AIDS Linked to the Intravenous Experience (ALIVE) study recruited injection drug users in Baltimore city and, to a small extent, surrounding counties.[27] While a wealth of studies has been conducted in the ALIVE Cohort, a particular study examining longitudinal injection drug use patterns is most relevant. In Figure 1.2, five distinct drug use patterns over time were shown (early cessation, delayed cessation, late cessation, frequent relapse and persistent injection).[2] Heavy drug use was also shown to be a barrier to HIV treatment initiation and adherence.[28] Drug use remains a major issue that undermines people's health in Baltimore.

Injection drug use is associated with substantial mortality and morbidity.[29,30] In

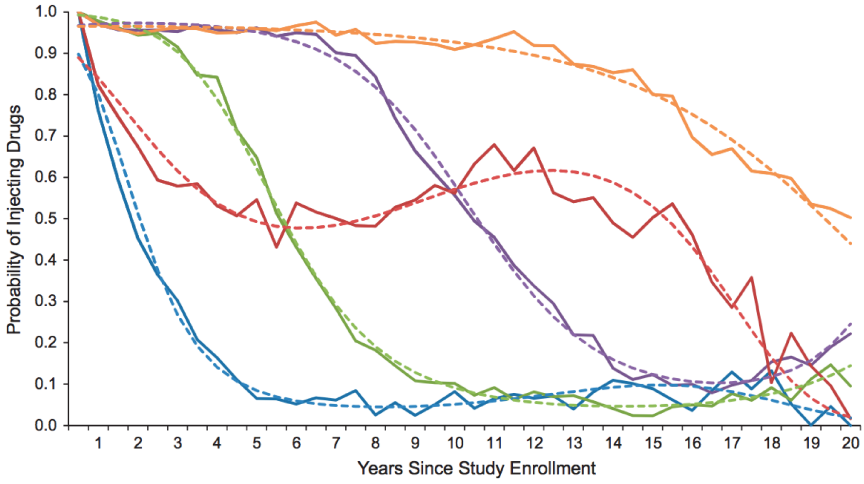


Figure 1. Trajectories of injection drug use among 1,716 injection drug users in the AIDS Linked to the Intravenous Experience (ALIVE) Study, Baltimore, Maryland, 1988–2008. The dotted lines represent the predicted probabilities of injection drug use conditional on membership in one of the 5 drug-use groups, while the solid lines represent the observed proportion of injection drug use given group membership. The *y*-axis represents the conditional probability of injection drug use, while the *x*-axis reflects time since study enrollment. The 5 groups (and prevalence of group within sample) are depicted with the following colors: blue, early cessation (19%); green, delayed cessation (16%); purple, late cessation (18%); red, frequent relapse (16%); orange, persistent injection (32%).

Am J Epidemiol. 2011;173(7):829–836

Figure 1.2: Trajactories of injection drug use over 20 years in Baltimore[2]

Rome, the mortality risk among drug users was about 15 times higher compared to population controls among men, and 38 times higher among women.[29] Among injection drug users in Baltimore, the estimated standardized mortality ratio remained elevated excluding HIV-related mortality.[31] Also, studies have shown that injection drug users have much higher risk of HIV infection[32] and worse HIV outcome during treatment initiation compared with non-IDUs. Notably, injection drug use may affect HIV progression,[33,34] and may lead to shortened disease-free survival time in HAART treatment.[34] Among frequently injected drugs, heroin may have immunosuppressive effects in human by affecting T lymphocyte functions and inhibiting T cell signaling.[35–37]

However, the exact biological impact by injection drug use or HIV on human

health are still unknown, and it is hard to develop a valid biomarker for recent (e.g., past six month) injection drug use. Genetics studies, especially epigenetics studies open a door to better understand the changes in genomic and physiologic profiles in response to injection drug use and HIV, and possibly as a biomarker for recent injection drug use.

1.1.3.2 Genetics and Epigenetics

Evidence has shown that injection drug use is associated with biological modification of the genome, i.e., epigenetics markers such as DNA methylation and histone modification.[38,39] For example, the human mu opioid receptor (OPRM1) gene is shown to be associated with opioid injection use.[40] Several studies have shown that methylation near OPRM1 gene is associated with opioid use disorder, or changes in methylation in animal models. However, these approach do not assess the entire epigenome for associations.[41,42]

Up to now, the epigenome-wide association studies have been primarily conducted in tobacco and alcohol use in humans.[42] Thus, it is critical to study mechanisms of how injection drug use affects human health by comprehensively examine the entire epigenome.

In HIV, recent studies have shown that there are epigenetic signals associated with HIV infection.[43,44] The top CpG signal associated gene NLRC5 in these studies have also presented in HIV integration sites.[45] There is also other evidence that

epigenetics may play a role in HIV latency.[46] It will be interesting to further study the epigenome with better coverage of CpG sites and with careful adjustment for cell type heterogeneity between HIV positive and negative samples.

1.1.4 Autism Spectrum Disorder(ASD)

1.1.4.1 Epidemiology

Autism Spectrum Disorder (ASD) includes a range of complex neurodevelopmental disabilities characterized by social and communication impairments, as well as restricted, repetitive, and stereotyped patterns of behavior according to DSM-5.[47] ASD affects 1 in 68 children in the US according to estimates from CDC's Autism and Developmental Disabilities Monitoring Network.[48] Studies in Asia, Europe, and North America have identified children with ASD with median of prevalence estimates of autism spectrum disorders of 62/10,000.[49] Compared with non-ASD-associated illnesses, ASD was associated with $3020 higher health care costs and $14.061 higher non-health care costs.[50] Most families reported high levels of burden following their child's diagnosis.[51] The forecast annual direct medical, direct non-medical, and productivity costs of ASD combined will be $461 billion (range $276-$1011 billion; 0.982-3.600 % of GDP) for 2025.[52] However, there is no medication that can cure Autism Spectrum Disorders, but several behavior interventions during early development can greatly improve autistic childrens social performance.[53] In order to better understand

the causes of ASD and develop effective medication, there is a great need for research in ASD etiology.

1.1.4.2 Genetics and Epigenetics

The heritability of Autism Spectrum Disorder (ASD) is estimated to be about 50%, suggesting genetic factors play a major role in its etiology.[54–56] Several genome-wide association studies (GWAS) have been carried out for ASD in hopes of identifying common genetic variants that are associated with risk.[12,57–59] However, these have thus far likely been underpowered, since most GWAS assess the whole genome consisting of millions of single-nucleotide polymorphisms (SNPs), and few consistent inherited common risk variants have been identified.

The identified common genetics variants, inherited and de novo rare variation, copy number variants, contribute not specific to Autism, but related with other neurodevelopmental disorders such as intellectual disability and fragile X syndrome.[55,60,61] Epigenetics evidence also suggests possible shared biological mechanism among neurodevelopmental disorders.[23]

Neurodevelopmental disorders are a group of conditions during the developmental period characterized by developmental deficits that produce impairments of personal, social, academic, or occupational functioning.[47] According to CDC, about 1 in 6 children in the United States had a developmental disability in 2006-2008.[62] Intellectual disability, one common type of neurodevelopmental disorders, often comorbid with

ASD.[63] Some neurodevelopmental disorders like fragile X syndrome, shared similar symptoms and risk genetic variants with ASD.[61]

Significant genetic variants across psychiatric disorders thus far have been shown to be enriched for SNPs that associate with expression levels, particularly in brain.[12,13] These expression-associated SNPs are often termed expression quantitative trait loci (eQTLs). Emerging studies on expression in the brain and brain eQTLs have been recently published, and their data are publicly available.[64,65] Hence, with existing information on brain eQTLs applied, the testing space for GWAS is reduced to only SNPs known to control brain expression levels, and thus efficiently increase the statistical power of GWAS and require smaller sample sizes.[66,67] The approach can be further limited to brain eQTLs that are associated in brain development or specific brain region. Head circumference and imaging studies of toddlers with autism have identified that deviant brain overgrowth majorly occurs during the first five years of life in children with autism, and functional abnormalities are majorly found in cortex.[68,69] Subsetting SNPs to specific developmental period and/or brain regions further reduce the testing space and may focus on more relevant and meaningful genetic variants associated with ASD.

1.2 Statement of Aims

1.2.1 Aim 1

Aim 1: Conduct epigenome-wide association analyses on injection drug use (any current injection drug use, cocaine injection, heroin injection, etc.) in the ALIVE cohort. Each substance will be assessed separately, and generalized estimating equations (GEE) will be used to accommodate the longitudinal study design. Apply the statistical method correlation motif to jointly assess the epigenetic profiles of all types of injection drug use in the ALIVE cohort.

1.2.2 Aim 2

Aim 2: Conduct the epigenome-wide association analyses on HIV infection status and HIV viral load in the ALIVE cohort, and assess how injection drug use status affects HIV epigenetic profiles.

1.2.3 Aim 3

Aim 3: Conduct genome-wide association analyses(GWAS) in Autism Spectrum Disorder in the SEED study, and further subset the GWAS SNPs to brain eSNPs and compare the shapes of QQ plots.

1.3 Public Health Significance

The high HIV prevalence among IDUs is a substantial public health challenge world-wide.[34,70,71] IDUs often have faster HIV progression rate and worse HIV treatment outcome compared with non-IDUs.[33,34,71] However, few studies have examined the underlying biological mechanisms by which injecting drugs like opioid affects human metabolism. To address the gaps in the literature, we propose a novel epigenome-wide association analysis to investigate the DNA methylation changes due to drug injection.

Epigenetics studies in humans is important because it enables discoveries. A large amount of opioid studies take advantage of mice models and experiments that are infeasible in humans due to constraints in tissue collection and ethical concerns.

Identifying the significant regions in the epigenome could lay a good foundation for further discovery on the exact biological pathways of opioid and cocaine injection.

It is also worth noting that this study focuses on minority populations since the majority of samples in the ALIVE cohort are African-Americans, which are commonly underrepresented in genetic research. Evidence has shown that the racial disparity among IDUs is striking, with higher HIV prevalence in blacks compared with whites.[72] However, few genetic or epigenetic studies have focused on this population to investigate the etiology. Our study aims to fill in this gap and contribute to the overall understanding of the etiology of drug addiction and HIV progression in humans.

Multiple studies have reported increased prevalence of ASD.[73–75] Despite the in-

creasing trend of diagnosis of ASD, the etiology of ASD remains unknown and no medication has been developed to fully cure ASD. Genes have long been known to play a major role in ASD and are believed to provide important information on the underlying biological mechanism of ASD. However, due to lack of statistical power to test across the whole genome, few significant genetic variants were discovered and no consistent variants were found. Given the challenges of identifying common genetic variants associated with ASD in only moderately sized samples, strategic subsetting of the genome into relevant regions likely to harbor risk is a rational approach to a practical problem. Identification of significant brain eQTL SNPs associated with ASD can shed immediate light on the potential functionality of the association and therefore improve understanding of ASD mechanisms, which may further improve our understanding of the connection between genes, brain structure and behavior.

Chapter 2

Methods

2.1 Data

2.1.1 The AIDS Linked to the IntraVenous Experience (ALIVE) Study

The ALIVE study is a prospective cohort study characterizing the incidence and natural history of HIV infection among injection drug users (IDUs) in Baltimore, MD.[27] DNA methylation were collected by Illumina Methylation EPIC bead chip and in January 2016 for about 800 individuals.

At study inception, HIV serologic screening of 2,960 IDUs identified a cohort of 668 seropositives followed biannually with interviews, exams and biospecimen collection.[27] In parallel, HIV seronegative subjects were followed through ALIVE-1

(as immunological controls, N 150) and the ALIVE-2 cohort (N 1000). Through ALIVE-2 screening, 328 HIV seroconverters have been identified and followed in the HIV-positive protocol. Additional recruitment in 1994-95, 1998, 2000 and 2005-08 replenished a total of 454 HIV-infected and 1208 uninfected IDUs. Follow-up experience by recruitment period indicates that 4% die and 7% are lost to follow-up annually.

At each 6-month study visit, participants undergo 1) standardized questionnaires (interviewer- and computer-administered); 2) a locator form to record participant address and contacts; 3) clinical exam, and 4) biological specimen collection for routine laboratory and repository. Records from all reported medical encounters are requested with standardized abstraction for clinical outcomes. Linkage to administrative, disease and mortality databases provide additional endpoints.

ALIVE remains predominantly African American and majority male with a median age exceeding 50 years. From 90% actively injecting at study entry, less than one-third of current HIV-infected participants are active injectors. Non-injection drug use (mainly crack), smoking (79%) and regular alcohol use (43%) are highly prevalent.

2.1.2 The Study to Explore Early Development (SEED) Study

The SEED study consists of 1309 samples, with 584 Autism Spectrum Disorder (ASD) cases and 725 normal controls with multiple ancestry, and over 30 million single-nucleotide polymorphisms (SNPs) that passed quality control (QC). The genotyping platform used was either the Illumina Omni1M Quad or the Affymetrix Kaiser Axiom array. The QC measures included removal of samples with low call rates (<98%), sex discrepancies, inappropriate relatedness (pi-hat > 0.2), and/or excess heterozygosity or homozygosity, removal of SNPs with a minor allele frequency less than (MAF<5%), and flagging of SNPs statistically significant ($p < 10^{-6}$) for not being in Hardy Weinberg Equilibrium in ancestry-stratified control samples. Following application of QC filters, phasing was performed using SHAPEIT followed by imputation against the 1000 Genome Project panel using IMPUTE2, resulting in over 30 million SNPs per sample. Ten principal components representing the multiple genetic ancestry of each sample were used to adjust for ancestry in GWAS analyses.

2.1.3 Psychiatric Genomics Consortium (PGC)

The Psychiatric GWAS consortium[76] aims to integrate genomic data on psychiatric disorders from different sources to conduct GWAS meta- and mega-analyses. The consortium greatly increases the power of GWAS analysis and has contributed

preliminary findings on the genetic and biological basis of psychiatric disorders like schizophrenia. The pioneer publication[12] on five psychiatric disorders including autism sparked further genomics and functional genomics analyses on psychiatric disorders. The PGC-AUT panel was reported to have 3,303 cases and 3,428 controls. The most recent PGC-AUT has 9 million imputed SNPs available for analysis. However, only summary data such as SNP reference ID, effect size and p-value is available to download; we currently do not have access to raw data.

2.2 Longitudinal analysis

Many longitudinal studies quantified epigenetics patterns over time.[77] To examine the change in epigenetics profiles across time, we need to account for repeated measurements on the same subjects in the longitudinal design. Researchers have previously reported longitudinal changes in candidate epigenetics sites[78] using mixed effect model with random intercept. Other statistical methods that account for longitudinal design includes generalized estimating equations(GEE) and variance components.[79]

When these longitudinal methods applied into epigenome-wide analyses, computational time to account for longitudinal design is much longer than candidate epigenetics sites approaches. The following sections described two methods that can perform epigenome-wide longitudinal analyses within hours.

2.2.1 Linear mixed effect model

A linear mixed effect model handles correlation on the dependent variable by introducing random effect components.[80] In a mixed effect model, fixed effect parameters measure population mean response(between subject variation) just like the conventional linear model, and the random effect parameters model the variability of response from the same individual(within subject variation). By specifying a random intercept in the mixed model, the baseline level of response is considered random and subject-specific, and in this way it handles the innate correlation within subject.

Mixed effect modeling is a popular method in longitudinal cohorts, but when applied to genome-wide studies, the traditional R package *lme4* takes about 1 hour to compute 10 CpG sites.[81] Several mixed effect model packages in GWAS have been developed to account for population structure.[82–85] However, the only R package that can transfer to epigenetics data is the *lrgpr* R package with specification of the genetic similarity matrix K to be the block identity matrix based on subject ID.[82] The results based on linear mixed models were generated using the lrgpr package.

2.2.2 Generalized Estimating Equations (GEE)

The Generalized Estimating Equation (GEE) is another statistical method that can account for a longitudinal design.[86] However, instead of modeling the covariance matrix between repeated measurements, the GEE model adopts the sandwich esti-

mator to approximate the caovariance matrix. It has properties that the estimation of the coefficient is robust even if the working covariance matrix is misspecified.[87] Since it does not directly estimate the covariance structure, the computational speed is much faster. The R implementation of these methods are *gee* and *geepack* packages.[88]

In general, the mixed effect model with random intercept and GEE model yield similar p-values under the same covariate specification. Figure 2.1 shows the p values from both models on detecting significant CpG sites associated with IDU.

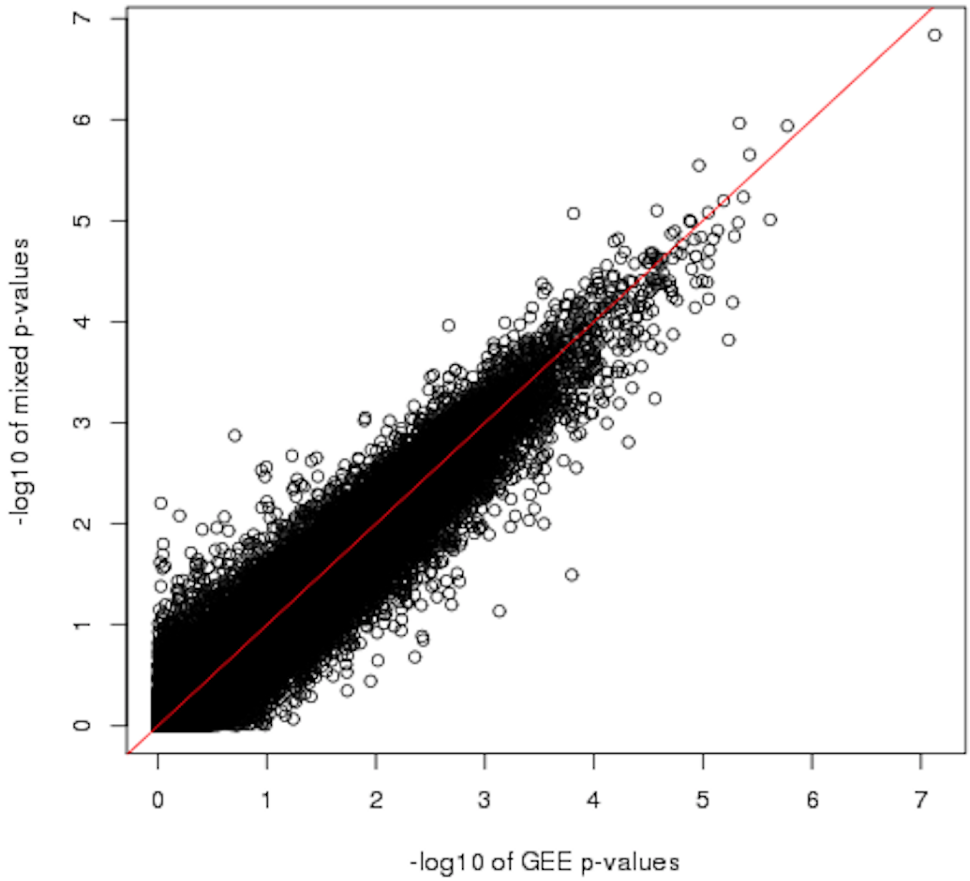


Figure 2.1: Concordance on p values between mixed effect model and GEE model (epigenome-wide scan of CpG sites on IDU)

2.3 Methods in Epigenetics

2.3.1 Preprocessing

2.3.1.1 Overview

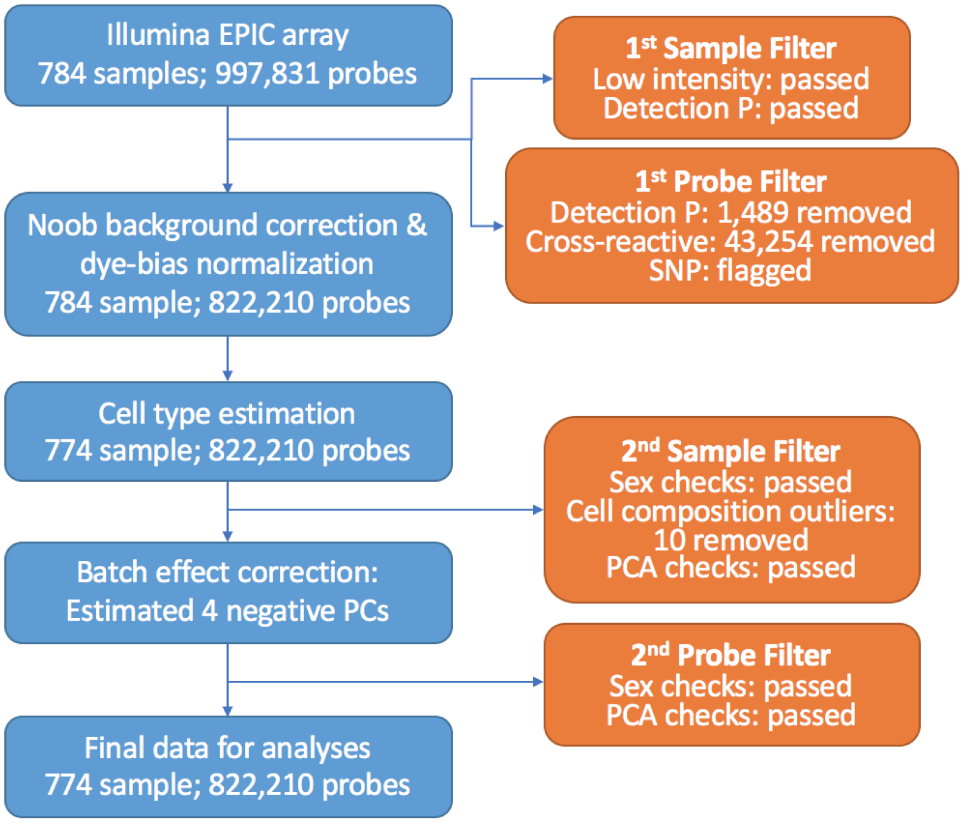


Figure 2.2: Overview of preprocessing pipeline

The Illumina methylation EPIC chip is designed to measure about 850,000 CpG methylation sites.[89] The raw red channel and green channel files generated from the chip need proper preprocessing and quality control before any analyses.[90] Major preprocessing steps include examining intensity between methylated and unmethy-

lated probes, density of beta and M value by probe type, PCA(principal component analysis) plots on the intensity values, background correction and normalization.[90] These quality control steps were conducted by using the R package *minfi*.[91,92] Figure 2.2 illustrated the preprocessing pipeline implemented in the ALIVE methylation analyses.

2.3.1.2 Intensity plots

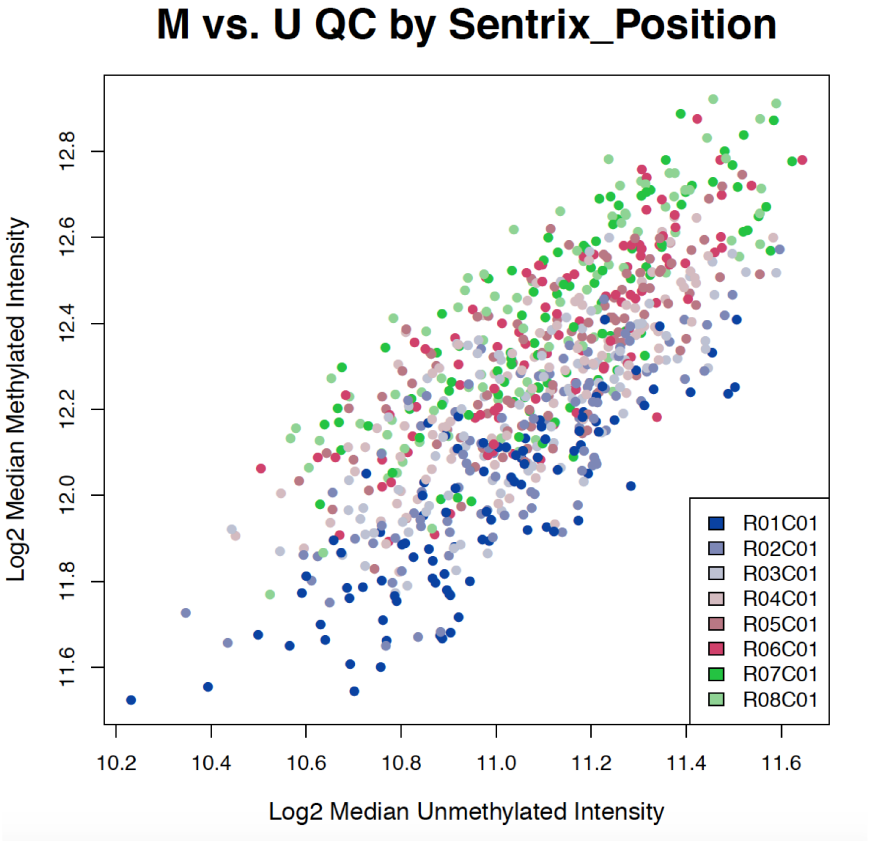


Figure 2.3: Methylated vs. Unmethylated Intensity

The Illumina assay utilizes two types of probes (methylated and unmethylated

probes) to detect the methylation level of a CpG site. The distribution of raw intensity values on methylated vs.unmethylated intensity values are first examined in Figure 2.3 colored by sentrix position. There is no obvious low intensity on either methylated or unmethylated intensity, demonstrating good quality.

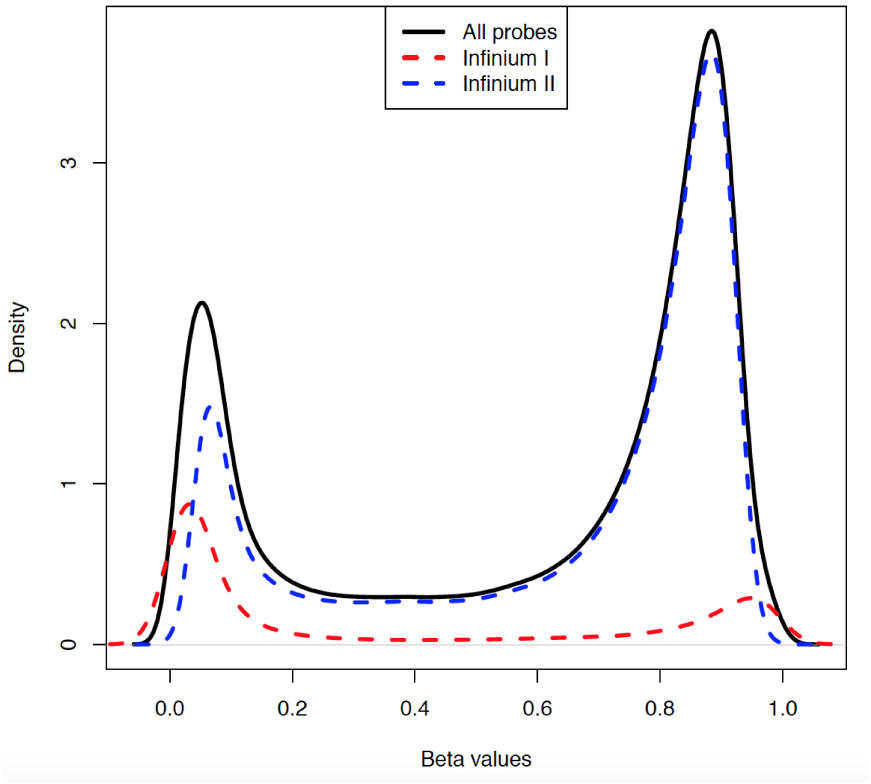


Figure 2.4: Raw beta intensity by probe type

A biologically meaningful measure of methylation level is the ratio of methylated intensity out of overall intensity, referred to as the beta value. Since the beta value is a ratio from 0 to 1, the alternative representation of methylation level is the logit

transformed version of beta value, referred to as the M value ranging from $-\infty$ to ∞.

$$\begin{aligned} \beta &= \frac{\text{Methylated intensity}}{\text{Methylated intensity} + \text{Unmethylated intensity} + 100} \\ M &= \log_2 \frac{\text{Methylated intensity} + 1}{\text{Unmethylated intensity} + 1} \end{aligned}$$

Beta values are often easier to interpret in terms of biological meaning, whereas the M value distribution is more normal and less variable.[93] Some researchers have argued that the M value is preferred in epigenetics analysis with better a detection rate and true positive rate.[93] The intensity of raw beta values by two Illumina probe types are shown in Figure 2.4.

2.3.1.3 Probe filter

In addition to low intensity probes, probe filters on detection p-value, cross reactive probes and probes that are single nucleotide polymorphisms (SNPs) are commonly used.

Detection p-value quantifies the probability of detecting the intensity from the background noise, not from the true signal.[90] Probes that have $p > 0.05$ are considered not likely to be the true signal and usually filtered out. The distribution of failed fraction of sample per probe by detection p-value is shown in Figure 2.5.

Some probes in the Illumina array are mapped to multiple parts of the genome, and are called cross-reactive probes.[94,95] These probes are usually removed from the

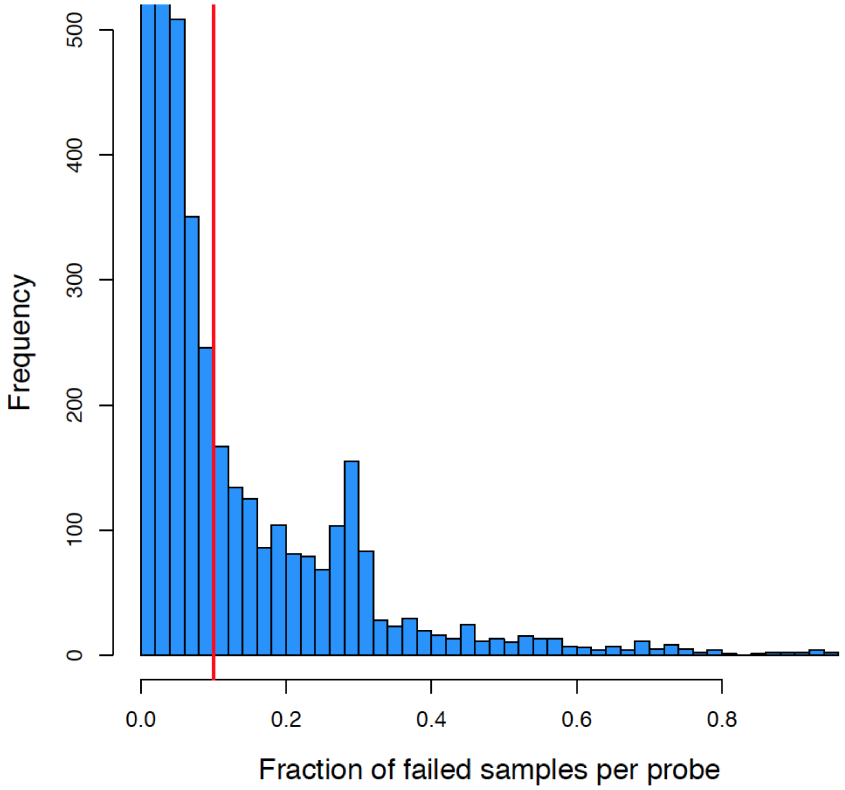


Figure 2.5: Fraction of failed sample per probe based on detection P value

analyses to ensure one-to-one mapping accuracy between the probes and their target CpG methylation sites.

Additionally, some probes target on SNPs in the genome, and SNP variation may affect methylation level in these CpG sites. These probes and CpG sites are usually flagged in the analysis.

2.3.1.4 Sample filter

The methylation differences on X chromosome and Y chromosome can be captured and *minfi* can estimate sex based on methylation profiles. If there is any mismatch

between the estimated sex and sex information from phenotype data, there may be quality control issues on the sample and may need to be removed. An estimated sex plot from *minfi* colored by actual reported sex is shown in Figure 2.6.

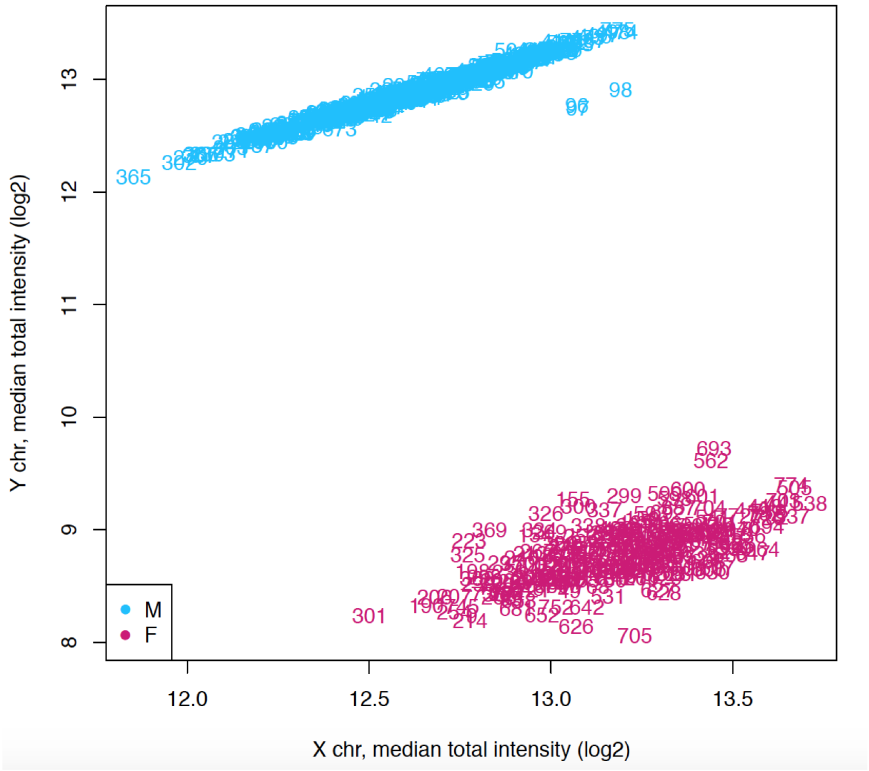


Figure 2.6: Median total methylation density by X and Y chromosome colored by actual sex

For white blood cells samples, cell composition can also be estimated based on the DNA methylation, including cell proportions of CD4+ and CD8+ T-cells, natural killer cells, monocytes, granulocytes, and B-cells.[96] Sample outliers with abnormal cell composition may also need to be removed.

2.3.1.5 Background correction and normalization

The noise from background in the Illumina platform often needs to be removed from the analysis, and background correction can be done by *noob* background correction (normal-exponential convolution using out-of-band probes) with dye-bias normalization to remove technical variation within sample.[90,91,97]

2.3.1.6 Batch effects adjustment

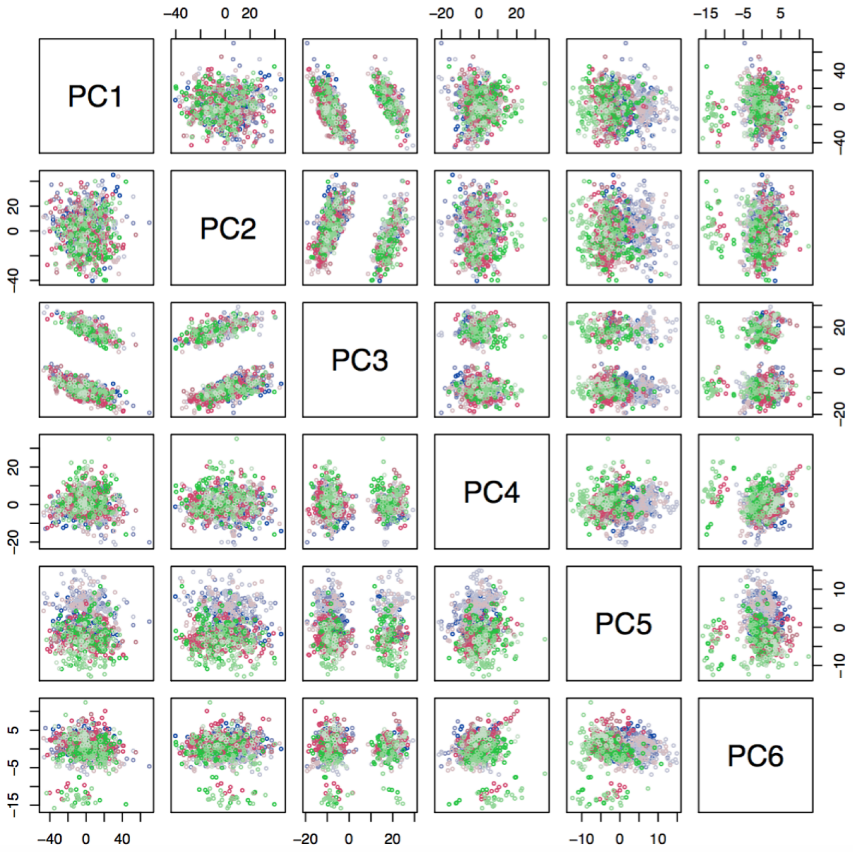


Figure 2.7: PCA of methylation intensity colored by plate reveal clusters in PC5, indicating batch effect

Batch effects are technical variations that may overshadow the true biological

signal. Methylation blood samples measured from the same plate, slide, slide position, or same time may cluster together due to batch effects. The PCA plot colored by plate in Figure 2.7 shows the probes from the same plate tend to cluster together and result in technical artifacts.

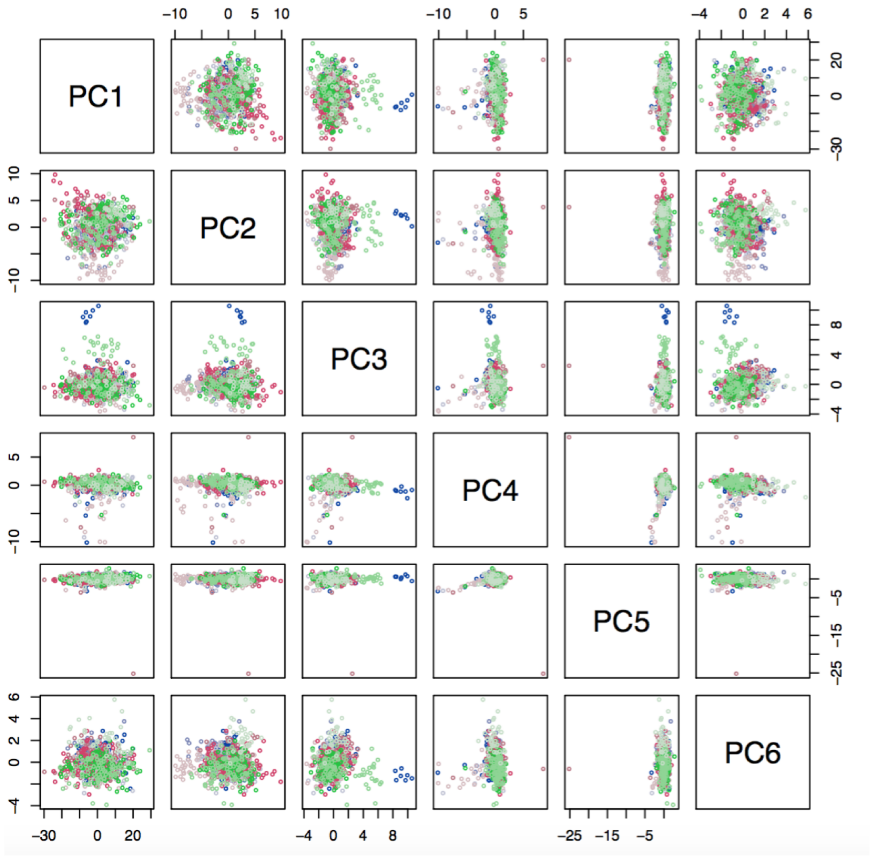


Figure 2.8: PCA of negative control PCs colored by plate reveal PC3 can capture plate effect

Many statistical methods have been established to correct for batch effects.[98–102] The method *combat* can correct for user specified known batch effect, and can create a new matrix of intensity after adjustment.[98] However, *combat* cannot adjust for unknown batch effects and cannot account for multiple sources of batch effect.

The SVA and ISVA methods can account for unknown batch effects,[99,100] but when implemented in our data, it generates more than 200 SVs and cannot be applied directly to longitudinal methylation data. The RUV method utilized negative control features that are only correlated with technical variations.[101] We can extract top PCs from those negative control probes and it can be directly applied to longitudinal data. Figure 2.8 shows the top negative control PCs that capture batch effect by plate.

2.3.2 Cell composition estimation

DNA methylation is greatly affected by cell type, and thus adjusting for cell composition is very important in DNA methylation related analyses.[103] The cell composition can be estimated either by reference-based methods or reference free methods.[96,104] The widely used reference based method is obtained by 46 blood samples from people of European ancestry,[96] and thus may not be as accurate when applying to samples of different ancestry or with infection such as HIV. The reference free approach gives flexibility to accommodate different sources of data, but the users need to specify number of clusters and the estimated variables are hard to interpret.[104] We obtained measured CD4+ and CD8+ cell proportions from about 200 blood samples and compare them with the estimated Houseman cell proportions in Figure 2.9, and it seems that the cell proportions are underestimated based on the Houseman method.

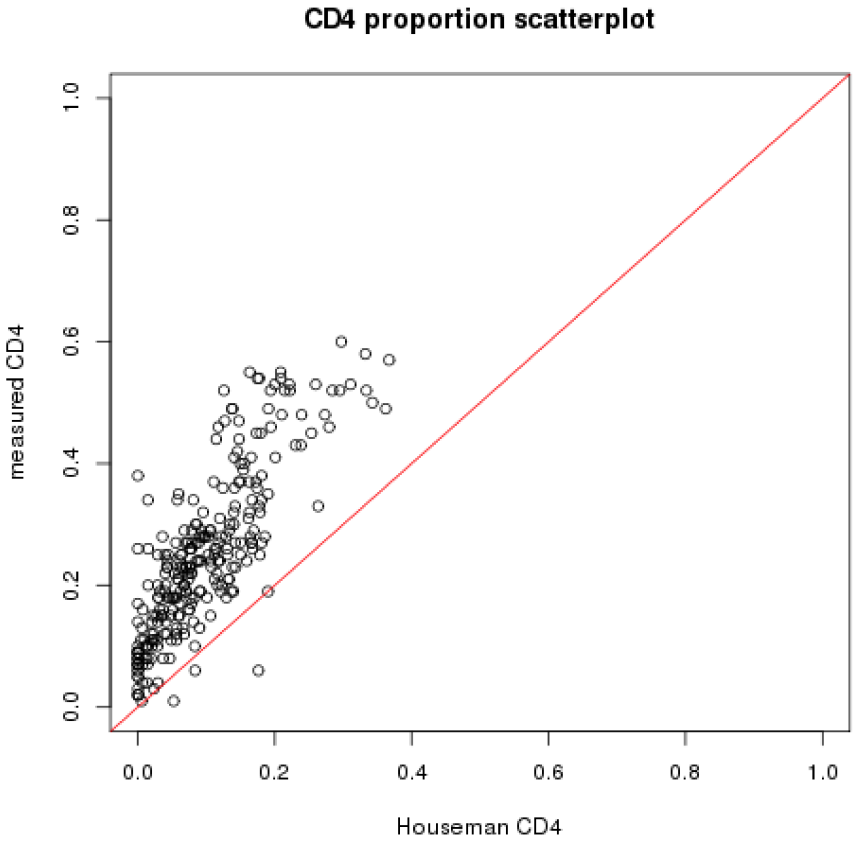


Figure 2.9: Estimated vs. measured CD4+ proportions

2.3.3 Single site analysis

Just like genome-wide association analysis, we can run site-by-site analyses to explore each CpG site's association with the phenotype. However, there are some blocks of CpG sites that are correlated with each other similar to linkage disequilibrium in GWAS, but have not been well-defined in DNA methylation analyses. Region based analysis is usually done when the sample size is small by bump hunting.[105]

2.3.4 Gene set enrichment analysis

Most CpG sites have been annotated by the nearest gene in *minfi*.[91] To examine what gene pathway is enriched in epigenome-wide association analyses, we can run gene set enrichment analyses among moderately significant CpG associated genes. We use existing gene pathways databases such as GO(Gene Ontology) and KEGG(Kyoto Encyclopedia of Genes and Genomes).[106,107]

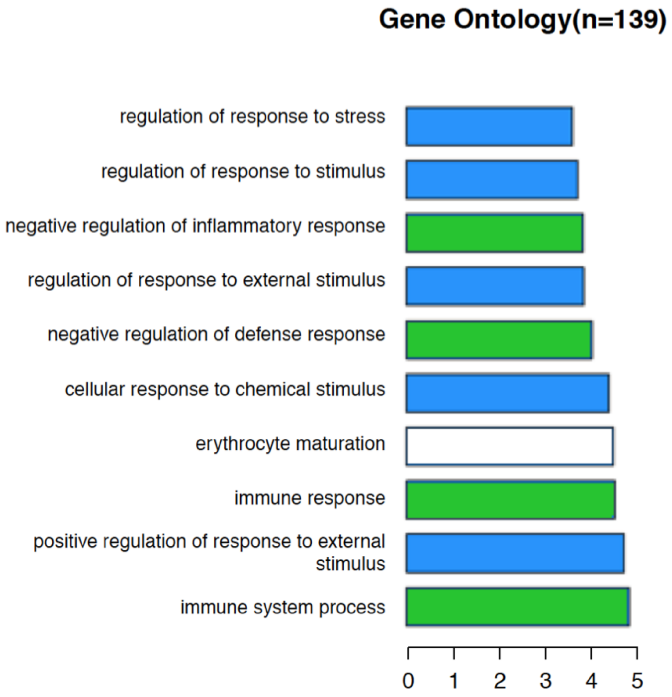


Figure 2.10: An example of gene set enrichment analysis, immune process(green) and response to external stimulus(blue) are enriched in significant CpG sites associated with HIV infection

It has been reported that there are potential biases since the probes in DNA methylation array are not sampled uniformly across the entire genome, and thus may give rise to bias that certain gene pathway are enriched as a result of sampling but

not a true association between phenotype and DNA methylation markers.[108] Thus, we used a gene set enrichment analysis with prior to correct for sampling bias in *missMethyl*.[109] An example of gene set enrichment is shown in Figure 2.10.

2.3.5 Epigenetic clock

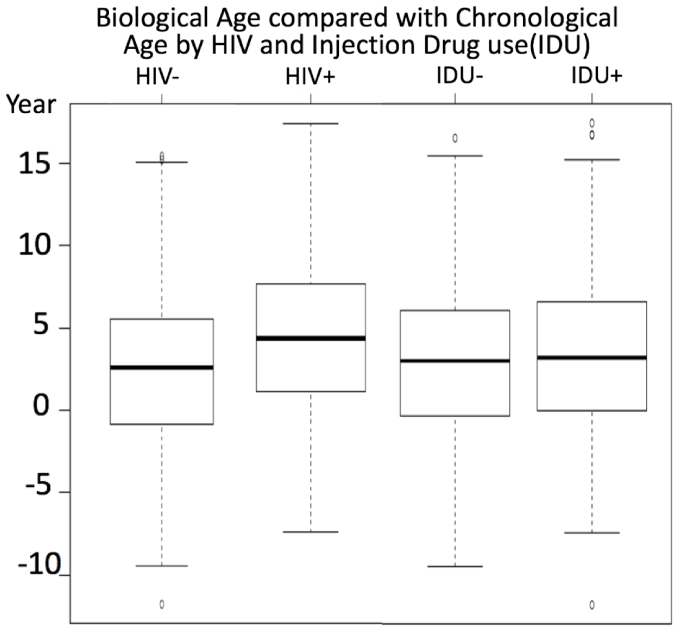


Figure 2.11: The difference between biological age estimated from DNA methylation data and chronological age, showing that HIV+ individuals age faster than HIV-, but there is no difference between current injection users and non-users

Several DNA methylation sites have been shown to be associated with the biological age, which might be different from chronological age and reflective of aging speed.[110] It is reported that the HIV infection is associated with accelerated speed of aging.[111] In our analyses, we also compared the difference between estimated biological age vs. chronological age by HIV infection status and injection drug use in

Figure 2.11.

2.4 Joint analysis by correlation motif

In any genetics and epigenetics studies, usually multiple phenotypic measures were collected for the sample. These phenotypes are usually correlated, for example, different types of drug use status for the past six months. Typically, we run separate genetics and epigenetics analyses on these phenotypes, ignoring the fact that there are potential correlation between the analyses. A joint analysis over these correlated phenotypes can be a better approach. By borrowing information across phenotypes, we may detect genetic or epigenetic variants that cannot be discovered by separate analyses alone. By using the summary level statistics, we want to examine if there are groups of genetic or epigenetic variants with the same patterns of association across the phenotypes. By accounting for the correlation across phenotypes, we can generate new statistics that may lead to new discovery and insight to the phenotype of interest.

There are existing methods in GWAS for jointly analyzing different phenotypes to detect significant SNPs. The method *ASSET* search through subsets of phenotypes to find the optimal subset for meta-analysis.[112] *MultiPhen* tests the association between SNP and the linear combination of phenotypes.[113] However, not many of these methods have been used in jointly analyzing epigenetic data like DNA methylation.

Thus, we want to explore if joint analysis in DNA methylation data will inform new epigenetic variant across phenotypes.

2.4.1 Correlation motif in expression data

Developed by Dr. Hongkai Ji's group, the joint analysis method called correlation motif can meta-analyze gene expression data in different studies.[114] An overview of the method is shown in Figure 3.6.

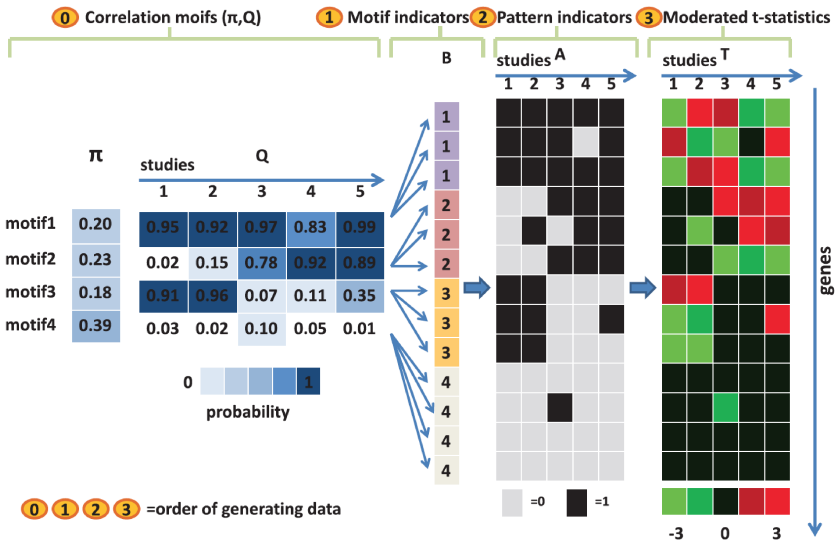


Figure 2.12: Illustration of correlation motif method

Let g indicates the number of genes and d indicates the number of studies in separate analyses. $\mathbf{T} = (t_{gd})$ is a $G \times D$ input matrix of summary test statistics from separate analysis. a_{gd} is the indicator on whether the gene a is deferentially expressed in study d. The proposed method first finds K groups of genes that have similar differential expression pattern across different studies with the group label as

b_g with probability $\pi = (\pi_1, ..., \pi_K)$, where $\pi_k = Pr(b_g = k)$ and $\sum_k \pi_k = 1$. The probability of any gene in class $b_g = k$ to be associated with study d is defined as $q_{kd} = Pr(a_{gd} = 1|b_g = k)$, which is the correlation structure between groups of genes and study. f_0 is the null distribution of t_{gd} and f_1 is the alternative distribution.

This method extracts correlation structure π_k, q_{kd} referred as a correlation motif, and then calculates the posterior probability based on the correlation motif as a prior and data from separate analyses. The joint probability is defined as follows:

$$Pr(\pi, \mathbf{Q}, \mathbf{A}, \mathbf{B}|\mathbf{T}) \propto \prod_{g=1}^{G}\prod_{k=1}^{K}\left\{\pi_k \prod_{d=1}^{D}\left[q_{kd} f_{d1}(t_{gd})\right]^{a_{gd}}\left[(1-q_{kd}) f_{d0}(t_{gd})\right]^{1-a_{gd}}\right\}^{\delta(b_g=k)} * \prod_{k=1}^{K}\pi_k \prod_{k=1}^{K}\prod_{d=1}^{D} q_{kd}(1-q_{kd})$$

where we use a Dirichlet prior for π and a beta prior $B(2, 2)$ for q_{kd}.

To determine the optimal number of group K, we use Bayesian information criterion(BIC) as the indicator for model selection. The output is whether each gene g is significant in study d, i.e., a_{gd}, and is by default determined by the posterior probability greater than 0.5.

2.4.2 Correlation motif applied in DNA methylation data

To make the correlation motif method applicable to different phenotypes in DNA methylation data, we made several changes to the model. First, we used p-values instead of test statistics as summary statistic input from separate analyses. Second, the null distribution f_0 is assumed to be uniform (0,1), and the alternative distribution f_1 follows a beta distribution estimated by permutation. Third, we calculate the local FDR using 1-posterior probability to ensure we can make comparisons between separate analyses and joint analyses. Details on how to formulate the model and run the EM algorithm are shown in the following sections.

2.4.2.1 E-step

The joint probability with priors are:

$$Pr(\pi, \mathbf{Q}, \mathbf{A}, \mathbf{B}|\mathbf{T}) \propto \prod_{g=1}^{G}\prod_{k=1}^{K}\left\{\pi_k \prod_{d=1}^{D}\left[q_{kd}f_{d1}(t_{gd})\right]^{a_{gd}}\left[(1-q_{kd})f_{d0}(t_{gd})\right]^{1-a_{gd}}\right\}^{\delta(b_g=k)} * \prod_{k=1}^{K}\pi_k \prod_{k=1}^{K}\prod_{d=1}^{D} q_{kd}(1-q_{kd})$$

The log-likelihood is:

$$\begin{aligned}
\ln Pr(\pi, \mathbf{Q}, \mathbf{A}, \mathbf{B}|\mathbf{T}) &= \sum_{g=1}^{G}\sum_{k=1}^{K}\delta(b_g = k)\ln \pi_k \\
&\quad + \sum_{g=1}^{G}\sum_{k=1}^{K}\delta(b_g = k)\{\sum_{d=1}^{D} a_g d[\ln q_{kd} + \ln f_{d1}(t_{gd})] \\
&\quad + \sum_{d=1}^{D}(1 - a_g d)[\ln(1 - q_{kd}) + \ln f_{d0}(t_{gd})]\} \\
&\quad + \sum_{k=1}^{K}\ln \pi_k + \sum_{k=1}^{K}\sum_{d=1}^{D}[\ln q_{kd} + \ln(1 - q_{kd})] + constant
\end{aligned}$$

Thus, the expectation of loglikehood with respect to $A, B|T, \hat{\pi}^{old}\hat{Q}^{old}$ is:

$$\begin{aligned}
Q(\pi, \mathbf{Q}|\hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}}) &= E_{old}\left[\ln Pr(\pi, \mathbf{Q}, \mathbf{A}, \mathbf{B}|\mathbf{T})\right] \\
&= \sum_{g=1}^{G}\sum_{k=1}^{K}\ln \pi_k E_{old}[\delta(b_g = k)] \\
&\quad + \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{d=1}^{D}[\ln q_{kd} + \ln f_{d1}(t_{gd})]E_{old}[\delta(b_g = k)a_{gd}] \\
&\quad + \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{d=1}^{D}[\ln(1 - q_{kd}) + \ln f_{d0}(t_{gd})]E_{old}[\delta(b_g = k)(1 - a_{gd})] \\
&\quad + \sum_{k=1}^{K}\ln \pi_k + \sum_{k=1}^{K}\sum_{d=1}^{D}[\ln q_{kd} + \ln(1 - q_{kd})] + constant
\end{aligned}$$

2.4.2.2 M-step

To obtain the maximum values after the E-step, we take the first order of partial derivative of $Q(\pi, \mathbf{Q}|\hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}})$ as follows:

$$\begin{aligned} \frac{\partial Q\left(\pi, \mathbf{Q}|\hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}}\right)}{\partial \pi_k} &= 0 \\ \frac{\partial Q\left(\pi, \mathbf{Q}|\hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}}\right)}{\partial q_{kd}} &= 0 \end{aligned}$$

We have:

$$\begin{aligned} \hat{\pi}_k^{new} &= \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1}{G + K} \\ \hat{q}_{kd}^{new} &= \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + 1}{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 2} \end{aligned}$$

Details on how to derive the maximum is shown below:

To obtain $\hat{\pi}_k^{new}$, since $\sum_{k=1}^{K} \pi_k = 1$ and $E_{old}\left[\delta(b_g = k)\right] = Pr_{old}(b_g = k)$, we have:

$$\begin{aligned}
\frac{\partial Q\left(\pi, \mathbf{Q} | \hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}}\right)}{\partial \pi_k} &= 0 \\
\frac{1}{\pi_k} - \frac{1}{\pi_K} + \frac{1}{\pi_k} \sum_{g=1}^{G} Pr_{old}(b_g = k) - \frac{1}{\pi_K} \sum_{g=1}^{G} Pr_{old}(b_g = K) &= 0 \\
\frac{1}{\pi_k}\left[\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1\right] &= \frac{1}{\pi_K}\left[\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1\right] \\
\pi_k &= \frac{\pi_K\left[\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1\right]}{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1}
\end{aligned}$$

Since $\sum_{k=1}^{K} \pi_k = 1$, then:

$$\begin{aligned}
\sum_{k=1}^{K} \pi_k = \frac{\sum_{k=1}^{K} \pi_K\left[\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1\right]}{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1} &= 1 \\
\frac{\pi_K\left[\sum_{g=1}^{G}\sum_{k=1}^{K} Pr_{old}(b_g = k) + \sum_{k=1}^{K} 1\right]}{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1} &= 1 \\
\frac{\pi_K\left[\sum_{g=1}^{G} 1 + K\right]}{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1} &= 1 \\
\frac{\pi_K(G + K)}{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1} &= 1 \\
\pi_K &= \frac{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1}{G + K}
\end{aligned}$$

Thus,

$$\hat{\pi}_k^{new} = \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1}{G + K}$$

To obtain $\hat{q}_{kd}^{new}$, we have

$$\begin{aligned} \frac{\partial Q\left(\pi, \mathbf{Q}|\hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}}\right)}{\partial \pi_k} &= 0 \\ \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + 1}{q_{kd}} - \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 0) + 1}{1 - q_{kd}} &= 0 \end{aligned}$$

Thus,

$$\begin{aligned} \hat{q}_{kd}^{new} &= \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + 1}{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + \sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 0) + 2} \\ &= \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + 1}{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 2} \end{aligned}$$

By definition of $Pr_{old}(b_g = k)$, it is the total probability of genes assigned to motif k.

Thus,

$$\begin{aligned} Pr_{old}(b_g = k) &= \frac{\hat{\pi}_k^{(old)} \prod_{d=1}^{D} [\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1 - \hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})]}{\sum_{l=1}^{K} \hat{\pi}_l^{(old)} \prod_{d=1}^{D} [\hat{q}_{ld}^{(old)} f_{d1}(t_{gd}) + (1 - \hat{q}_{ld}^{(old)}) f_{d0}(t_{gd})]} \\ Pr_{old}(b_g = k, a_{gd} = 1) &= Pr_{old}(a_{gd} = 1 | b_g = k) * Pr_{old}(b_g = k) \\ &= \frac{\hat{q}_{kd}^{(old)} f_{d1}(t_{gd})}{\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1 - \hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})} * Pr_{old}(b_g = k) \end{aligned}$$

The posterior distribution is given by:

$$\begin{aligned} E(a_{gd}|\mathbf{T},\hat{\pi},\hat{\mathbf{Q}}) &= Pr(a_{gd}=1|\mathbf{T},\hat{\pi},\hat{\mathbf{Q}}) \\ &= \int_k Pr(b_g=k, a_{gd}=1) \\ &= \sum_{k=1}^{K} \frac{\hat{q}_{kd}^{(old)} f_{d1}(t_{gd})}{\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1-\hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})} * Pr_{old}(b_g=k) \end{aligned}$$

According to how we set up the problem, the log-likelihood is:

$$\begin{aligned} \ln Pr(\mathbf{T}|\pi,\mathbf{Q}) &= \sum_{g=1}^{G} \ln \left\{ \sum_{k=1}^{K} \{\hat{\pi}_k^{(old)} \prod_{d=1}^{D} [\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1-\hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})]\} \right\} \\ &\quad + \sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K} \sum_{d=1}^{D} [\ln q_{kd} + \ln(1-q_{kd})] \end{aligned}$$

The BIC for K is:

$$\begin{aligned} BIC(K) &= -2 \sum_{g=1}^{G} \ln Pr(\mathbf{T}|\pi,\mathbf{Q}) + (K-1+K*D) * \ln G \\ &= -2 \ln \left\{ \sum_{k=1}^{K} \{\hat{\pi}_k^{(old)} \prod_{d=1}^{D} [\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1-\hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})]\} \right\} \\ &\quad + (K-1+K*D) * \ln G \end{aligned}$$

For missing data in phenotype d and CpG g, we have $f_{d1}(t_{gd}) = 1$ and $f_{d0}(t_{gd}) = 1$.

In $Pr_{old}(b_g = k)$, the numerator becomes:

$$\hat{\pi}_k^{(old)} [\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1-\hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})] = \hat{\pi}_k^{(old)} [\hat{q}_{kd}^{(old)} + (1-\hat{q}_{kd}^{(old)})] = \hat{\pi}_k^{(old)}$$

In $Pr_{old}(b_g = k, a_{gd} = 1)$:

$$\begin{aligned} Pr_{old}(b_g = k, a_{gd} = 1) &= \frac{\hat{q}_{kd}^{(old)} f_{d1}(t_{gd})}{\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1 - \hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})} * Pr_{old}(b_g = k) \\ &= \hat{q}_{kd} * Pr_{old}(b_g = k) \end{aligned}$$

Thus, we can substitute $f_{d0}(t_{gd}), f_{d1}(t_{gd})$ with 1 for the missing observations, and it does not affect the final estimation. We also use the following formula to make sure a and b are close to zero, the estimation process will remain accurate.

$$\begin{aligned} \log(a+b) &= \log(\max(a,b)\left(\frac{a}{\max} + \frac{b}{\max}\right) = \log\max + \log\left(\frac{a}{\max} + \frac{b}{\max}\right) \\ &= \log\max + \log\left(e^{\log a - \log\max} + e^{\log b - \log\max}\right) \end{aligned}$$

2.4.2.3 Implementation of E-M algorithm

1. Choose initial values for $\pi_k^{(0)}$ and $q_{kd}^{(0)}$.

2. Calculate $Q^{(1)}(\pi, \mathbf{Q}|\hat{\pi}^{(\mathbf{0})}, \hat{\mathbf{Q}}^{(\mathbf{0})})$ based on $\pi_k^{(0)}$ and $q_{kd}^{(0)}$:

(1) $a_{1,kdg} = \ln[\hat{q}_{kd}^{(old)} f_{d1}(t_{gd})]$; for missing data, $a_{1,kdg} = \ln \hat{q}_{kd}^{(old)}$ (3d-array)

$a_{2,kdg} = \ln[(1 - \hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})]$; for missing data, $a_{2,kdg} = \ln(1 - \hat{q}_{kd}^{(old)})$ (3d-array)

(2) $a_{max,kdg} = pmax(a_{1,kdg}, a_{2,kdg})$ (3d-array)

$e_{1,kdg} = e^{a_{1,kdg} - a_{max,kdg}}$ (3d-array)

$e_{2,kdg} = e^{a_{2,kdg} - a_{max,kdg}}$ (3d-array)

(3) $logp_{kg} = \ln \hat{\pi}_k^{(old)} + \sum_{d=1}^{D} \{\ln[\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1 - \hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})]\}$

$$= \ln \hat{\pi}_k^{(old)} + \sum_{d=1}^{D} \{\ln[e^{a_{1,kdg}} + e^{a_{2,kdg}}]\}$$

$$= \ln \hat{\pi}_k^{(old)} + \sum_{d=1}^{D} \{\ln[e^{a_{1,kdg}-a_{max,kdg}} + e^{a_{2,kdg}-a_{max,kdg}}] + a_{max,kdg}\}$$

$$= \ln \hat{\pi}_k^{(old)} + \sum_{d=1}^{D} [\ln(e_{1,kdg} + e_{2,kdg}) + a_{max,kdg}] \text{ (2d-array)}$$

$logpmax_{kg} = pmax(logp_{kg})$ (2d-array)

(4) $Pr_{old}(b_g = k) = \dfrac{e^{logp_{kg}-logpmax_{kg}}}{\sum_{l=1}^{K} e^{logp_{lg}-logpmax_{lg}}}$ (probability of a specific g belongs to motif k, output, 2d-array)

(5) $Pr_{old}(b_g = k, a_{gd} = 1) = \dfrac{e_{1,kdg}}{e_{1,kd} + e_{2,kdg}} * Pr_{old}(b_g = k)$; (3d-array)

$Pr_{old}(b_g = k, a_{gd} = 0) = \dfrac{e_{2,kdg}}{e_{1,kd} + e_{2,kdg}} * Pr_{old}(b_g = k)$; (3d-array)

(6) $Q^{(1)}(\pi, \mathbf{Q}|\hat{\pi}^{(\mathbf{0})}, \hat{\mathbf{Q}}^{(\mathbf{0})}) =$

$$\sum_{k=1}^{K} \ln \pi_k^{(0)} \left(\sum_{g=1}^{G} Pr_{old}^{(0)}(b_g = k) + 1 \right)$$

$$+ \sum_{g=1}^{G} \sum_{k=1}^{K} \sum_{d=1}^{D} [a_{1,kdg} Pr_{old}^{(0)}(b_g = k, a_{gd} = 1) + a_{2,kdg} Pr_{old}^{(0)}(b_g = k, a_{gd} = 0)]$$

$$+ \sum_{k=1}^{K} \sum_{d=1}^{D} [\ln q_{kd}^{(0)} + \ln(1 - q_{kd}^{(0)})] + constant$$

3. Calculate $\pi_k^{(1)}$ and $q_{kd}^{(1)}$ based on $Q^{(1)}(\pi, \mathbf{Q}|\hat{\pi}^{(\mathbf{0})}, \hat{\mathbf{Q}}^{(\mathbf{0})})$

$$\pi_k^{(1)} = \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1}{G + K} \text{ (1d-array)}$$

$$q_{kd}^{(1)} = \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + 1}{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 2} \text{ (2d-array)}$$

4. Repeat 2-3 for a maximum of n times, or when none of the parameters in π and $\mathbf{Q}$ changes by more than 0.1%.

5. The posterior distribution is: $E(a_{gd}|\mathbf{T}, \hat{\pi}, \hat{\mathbf{Q}}) = \sum_{k=1}^{K} \dfrac{e_{1,kdg}}{e_{1,kd} + e_{2,kdg}} * Pr_{old}(b_g = k)$

6. The marginal log-likelihood is:

$$\begin{aligned}
\ln Pr(\mathbf{T}|\pi,\mathbf{Q}) &= \sum_{g=1}^{G}\ln\left\{\sum_{k=1}^{K}\{\hat{\pi}_k^{(old)}\prod_{d=1}^{D}[\hat{q}_{kd}^{(old)}f_{d1}(t_{gd})+(1-\hat{q}_{kd}^{(old)})f_{d0}(t_{gd})]\}\right\} \\
&\quad +\sum_{k=1}^{K}\ln\pi_k+\sum_{k=1}^{K}\sum_{d=1}^{D}[\ln q_{kd}+\ln(1-q_{kd})] \\
&= \sum_{g=1}^{G}\ln\left\{\sum_{k=1}^{K}e^{logp_{kg}}\right\}+\sum_{k=1}^{K}\ln\pi_k+\sum_{k=1}^{K}\sum_{d=1}^{D}[\ln q_{kd}+\ln(1-q_{kd})] \\
&= \sum_{g=1}^{G}\left\{logpmax_{kg}+\ln\sum_{k=1}^{K}e^{logp_{kg}-logpmax_{kg}}\right\}+\sum_{k=1}^{K}\ln\pi_k \\
&\quad +\sum_{k=1}^{K}\sum_{d=1}^{D}[\ln q_{kd}+\ln(1-q_{kd})]
\end{aligned}$$

7. The BIC for K is:

$$BIC(K)=-2\sum_{g=1}^{G}\ln Pr(\mathbf{T}|\pi,\mathbf{Q})+(K-1+K*D)*\ln G$$

2.4.2.4 EM algorithm for estimating f_0, f_1

Set up:

1. Observed empirical p-values from permutation $\mathbf{X}$.

2. Missing probability of belonging to null distribution for each CpG site $\mathbf{Z}$.

3. Parameters to be estimated as listed below:

$f=\pi_0 f_0+\pi_1 f_1$

(1) f is the overall distribution.

(2) π_0 is the probability of being in the null distribution (expected to be > 0.99); needs to be estimated.

(3) π_1 is the probability of being in the alternative distribution (expected to be < 0.01); needs to be estimated.

(4) f_0 is the null distribution with Unif(0,1)

(5) f_1 is the alternative distribution with Beta(α,β) ; α, β needs to be estimated.

Steps:

1. Choose initial values for $\pi_0, \pi_1, \alpha, \beta$.

$\pi_0^{(0)} = 0.99, \pi_1^{(0)} = 0.01, \alpha^{(0)} = 1, \beta^{(0)}$=30

2. Compute f based on initial values.

3. Compute posterior probability for each CpG site belonging to f_0 as $\mathbf{Z}$.

Denote the model based p-value for CpG_i is p_i:

$$Z_i = \frac{\pi_0^{(0)} f_0(p_i)}{\pi_0^{(0)} f_0(p_i) + \pi_1^{(0)} f_1(p_i)}$$

4. Use $\mathbf{Z}$ to estimate $\pi_0^{(1)}, \pi_1^{(1)}, \alpha^{(1)}, \beta^{(1)}$.

$$\pi_0^{(1)} = \frac{\sum_i Z_i}{n},\ \pi_1^{(1)} = 1 - \frac{\sum_i Z_i}{n}$$

$$\alpha^{(1)} = \frac{\mu^2(1-\mu)}{\sigma^2} - \mu,$$

$$\beta^{(1)} = \frac{\mu(1-\mu)^2}{\sigma^2} + \mu - 1,$$

$$w_i = (1 - Z_i) / \sum_j (1 - Z_j)$$

$$\hat{\mu} = \sum_i w_i p_i$$

$$\hat{\sigma}^2 = \sum_i w_i (p_i - \hat{\mu})^2$$

5. Repeat 3-4 until converge.

Chapter 3

Epigenetics and Injection Drug Use

3.1 Introduction

There are estimated to be 11.0-21.2 million injection drug users (IDUs) world-wide (estimate from 2007), with HIV prevalence over 40% among IDUs in nine countries.[24] The prevalence of substance use is high in the US, with prevalence of 13.8% in 2012-2014 in large metropolitan areas in national survey reported past-month use of any illicit substance use.[25] According to the most recent results on National Survey of Drug Use and Health, there are about 475,000 people aged 12 or older were injecting heroin in 2016 in the US.[26] Estimates of cocaine use in 2016 were 1.9 million people in the US.

The exact biological mechanisms of the impact of injection drug use on human health are still unknown, and it is has been challenging to develop a valid biomarker

for recent injection drug use given the difficulties in accessing subjects and samples. Genetics studies, especially epigenetics studies open a door to better understand the changes in epigenetic profiles in response to injection drug use, and possibly as a biomarker for recent injection drug use.

Injection drug use is associated with substantial mortality and morbidity.[29,30] In Rome, the mortality risk among drug users was about 15 times higher compared to population controls among men, and 38 times higher among women.[29] Among injection drug users in Baltimore, the estimated standardized mortality ratio remained elevated excluding HIV-related mortality.[31] Also, studies have shown that injection drug users have much higher risk of HIV infection[32] and worse HIV outcomes during treatment initiation compared with non-IDUs. Notably, injection drug use may affect HIV progression,[33,34] and may lead to shortened disease-free survival time in HAART treatment.[34] Among frequently injected drugs, heroin is known to have immunosuppressive effects in human by affecting T lymphocyte functions and inhibiting T cell signaling.[35–37] Thus, it is important to examine how injection drug use impacts HIV latency and progression.

3.2 Method

The ALIVE study is an on-going prospective cohort study characterizing the incidence and natural history of HIV infection among injection drug users (IDUs) in

Baltimore, MD from 1988.[27] At each 6-month visit, clinical, behavioral and laboratory data such as HIV infection status and injection drug use were assessed for the participants. Blood was obtained from 288 current IDUs, resampled after cessation and then again after relapse (total samples = 774). HIV status did not change across visits.

The study participants went through HIV serology screening at baseline. HIV serology status was assessed at each study visit for HIV negative participants, whereas CD4+, CD8+ count and HIV viral load testing were performed at each study visit for HIV positive participants. Past six months injection drug use status and type, smoking patterns, and ART use information were obtained by computer-administered standardized questionnaires. The study design has been described in other literature.[27,115,116]

3.2.1 DNA methylation measurement and preprocessing

Peripheral blood mononuclear cell DNA was isolated with the Qiagen DNeasy kit and bisulfite converted with Zymo EZ methylation gold kit at the Johns Hopkins University Center for Inherited Disease Research. Bisulfite treated DNA was run on the Illumina Infinium MethylationEPIC BeadChip.

The *minfi* package (Bioconductor) was used to process raw Illumina image files

into noob preprocessed methylation beta values.[91] Cell composition on CD4+, CD8+, natural killer cells, monocytes, granulocytes and B cells were estimated based on the method described in Houseman et al.[96] and implemented in *minfi*.[91,103] Samples with low intensity, inconsistency on predicted and observed sex, and outliers in estimated cell composition were removed for quality control. Probes with low intensity or that are known to cross-hybridize were excluded. 774 samples and 822,210 probes were used in the final analyses. Batch effect was adjusted by top 4 PCs from negative control features that are only correlated with technical variations.[101] The pipeline of proprecessing is in supplementary figure 3.5.

3.2.2 Statistical analysis

3.2.2.1 Separate analyses using linear mixed effect model

M value is used in the analysis since the M-value distribution is closer to normal assumption and less variable. Researchers reported that the use of M value usually leads to better detection rates and true positive rates compared with the beta value.[93]

We used a linear mixed effect model with random intercept to account for the longitudinal design by using the R package *lrgpr*.[82] The package is designed for fast computation of mixed effect models to account for population stratification based on GWAS data,[82] but it allows flexibility on input data and the ability to manually specify the variance-covariance matrix. Thus, we can directly use this package to run

a linear mixed effect model with random intercept on epigenome-wide association analyses by specifying a genetic similarity matrix K as follows:

$$
\begin{aligned}
K &= [k_{ij}], \\
k_{ij} &= \begin{cases} 1, & \text{if } i,j \text{ are from the same subject} \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

where K is a $n \times n$ matrix and n is the total sample size.

We conducted single site association analyses on heroin only, cocaine only and heroin and cocaine co-use across the epigenome for the 822,210 CpG sites, adjusting for HIV status, gender, race, age, smoking status, cell composition and PCs for batch effect. The model is stated below(i indicates subject ID and t indicate time points):

$$
\begin{aligned}
M_{it} = {} & u_i + \beta_0 + \beta_1 \text{heroin only}_{it} + \beta_2 \text{cocaine only}_{it} + \beta_3 \text{co-use}_{it} + \beta_4 \text{HIV}_{it} + \beta_5 \text{gender}_{it} \\
& + \beta_6 \text{race}_{it} + \beta_7 \text{age}_{it} + \beta_8 \text{smoking}_{it} + \beta_9 \text{CD4}_{it} + \beta_{10} \text{CD8}_{it} \\
& + \beta_{11} \text{natural killer cells}_{it} + \beta_{12} \text{monocytes}_{it} + \beta_{13} \text{granulocytes}_{it} + \beta_{14} \text{B cells}_{it} \\
& + \beta_{15} \text{negative control PC1}_{it} + \beta_{16} \text{negative control PC2}_{it} + \\
& \beta_{17} \text{negative control PC3}_{it} + \beta_{18} \text{negative control PC4}_{it}
\end{aligned}
$$

For the analysis of any injection drug use, we replaced heroin only, cocaine only and co-use indicators with the overall indicator of any injection drug use. There are

108 visits that are missing specific injected drug type, and we imputed the drug type based on individual's past injection drug types and age.

$$
\begin{aligned}
M_{it} \;=\; & u_{it} + \beta_0 + \beta_1 \text{injection drug use}_{it} + \beta_2 \text{HIV}_{it} + \beta_3 \text{gender}_{it} \\
& + \beta_4 \text{race}_{it} + \beta_5 \text{age}_{it} + \beta_6 \text{smoking}_{it} + \beta_7 \text{CD4}_{it} + \beta_8 \text{CD8}_{it} \\
& + \beta_9 \text{natural killer cells}_{it} + \beta_{10} \text{monocytes}_{it} + \beta_{11} \text{granulocytes}_{it} + \beta_{12} \text{B cells}_{it} \\
& + \beta_{13} \text{negative control PC1}_{it} + \beta_{14} \text{negative control PC2}_{it} + \\
& \beta_{15} \text{negative control PC3}_{it} + \beta_{16} \text{negative control PC4}_{it}
\end{aligned}
$$

3.2.2.2 Joint analysis

After running separate epigenome-wide association analyses on any injection drug use, heroin only, cocaine only and co-use, we extracted the p-values from 4 analyses for each CpG sites and conduct joint analysis by correlation motif.[114] We estimated the correlation structure between groups of CpG sites and four phenotypes (any injection drug use, heroin only, cocaine only and co-use) by E-M algorithm, and calculated the posterior probability and local FDR incorporating the correlation structure. The number of CpG group is determined by the model with lowest bayesian information criterion(BIC). Details on the correlation motif method and its implementation in epigenome-wide association analyses can be found in supplementary methods and

Wei et al. 2014.[114] We also acknowledge that the original method is developed under conditional independence assumption, and any injection drug use is the sum of heroin only, cocaine only and co-use of heroin and cocaine. Thus, the results from the model should be interpreted carefully.

3.2.2.3 Gene ontology

The gene ontology enrichment analysis is done using in GO(Gene Ontology) database[106] by the R package *missMethyl*, with prior to correction for sampling bias.[108,109]

3.3 Results

3.3.1 Demographics

The sample characteristics are shown in Table 3.1. Among 288 subjects, about 2-4 visits per subjects were selected for methylation measurement illustrated in Figure 3.1. About half of visits were injection drug use visits for the past 6 months, and the other half the subjects did not inject any drugs. About 2/3 of the visits were male, and 1/3 of the visits were HIV positive. 92% of the samples were from African Americans. The average age of the sample visits was 49.

Table 3.1: Demographics of the ALIVE DNA methylation sample

Table 1: Study observation characteristics (n=774)		
	Mean(se)	N(%)
Average age	49.1±7.1	
Average visits	2.7±0.7	
Gender		
Male	531	68.6%
Female	243	31.4%
Race		
Caucasian/non-hispanic	39	5.0%
Caucasian/hispanic	0	0.0%
African Americans/non-hispanic	712	92.0%
African Americans/hispanic	8	1.0%
Asian	0	0.0%
Other	15	1.9%
Current Injection user		
Yes	379	49.0%
No	395	51.0%
HIV status		
Positive	272	35.1%
Negative	502	64.9%

3.3.2 Separate analysis

We ran linear mixed effect models accounting for the repeated measurement on the same subject. The QQ plots of any past six month injection drug use, heroin only injection, cocaine only injection and co-use of heroin and cocaine injection were obtained and showed no inflation(λ=0.96-1) in Figure 3.2.

DNA methylation at individual loci is significantly associated with cocaine and co-use of heroin and cocaine injection use after correction for multiple testing (Table 3.2). CpG sites associated with the FKBP5, AIM2 genes were ranked highest for

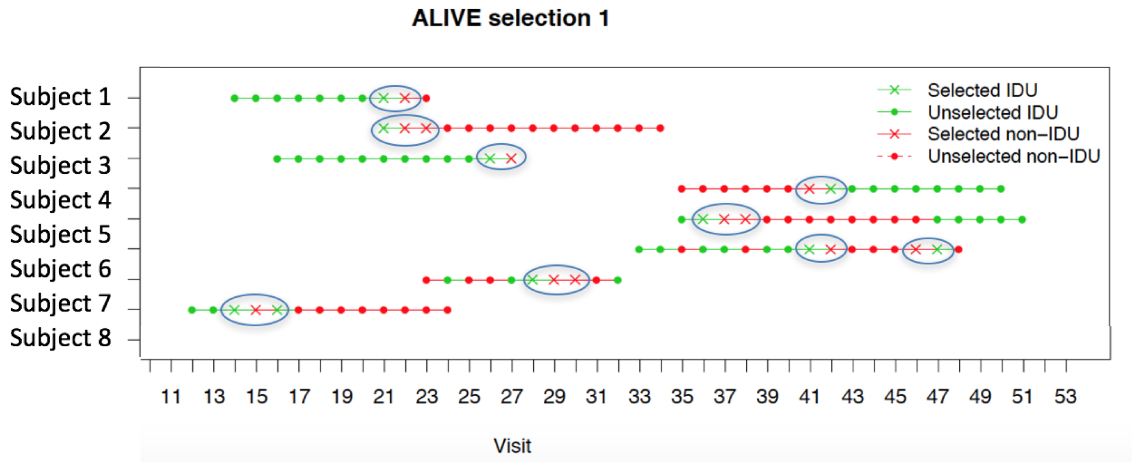


Figure 3.1: Illustration of ALIVE DNA methylation study design; subject's selected visits as circled include injection(green) and cessation(red)

association with IDU status.

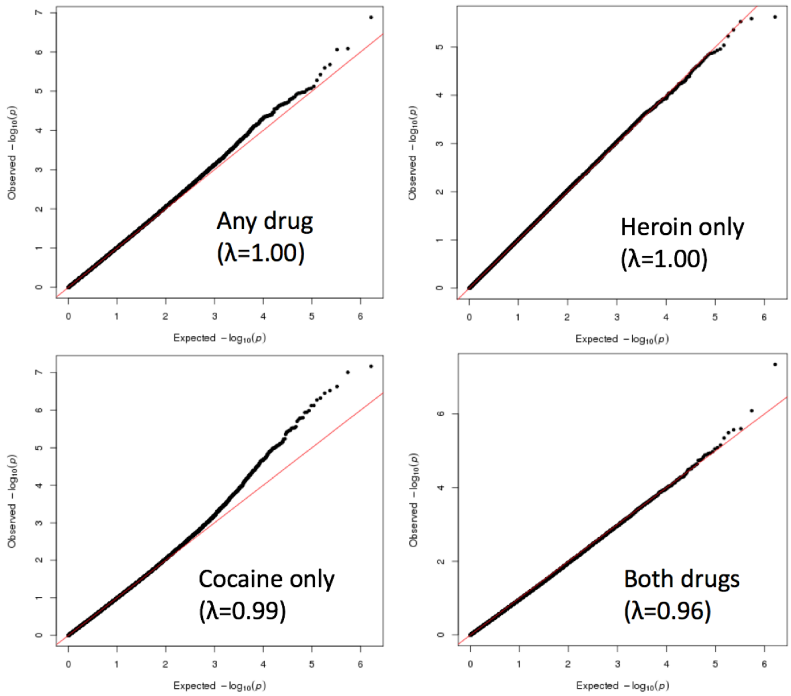


Figure 3.2: QQ plots of separate analyses on any injection drug use,

Table 3.2: Top hits in separate analyses in injection drug use

Table 2: Top 5 CpG sites associated with each four phenotype in separate analyses

Phenotype	Probe	CHR	BP	Gene	Beta	p-value	BH FDR
Any drug Use	cg03012169	chr9	110495498		-0.10	1.3E-07	1.1E-01
	cg03546163	chr6	35654363	FKBP5	-0.14	8.2E-07	2.4E-01
	cg03067296	chr17	76274577	AC087645.1	-0.07	8.7E-07	2.4E-01
	cg10636246	chr1	159046973	AIM2	-0.15	2.1E-06	4.2E-01
	cg10564101	chr8	101508599	KB-1615E4.2	0.08	2.6E-06	4.2E-01
Heroin only	cg15989617	chr14	61574389		-0.11	2.4E-06	8.2E-01
	cg22209773	chr14	65878437	FUT8	-0.10	2.6E-06	8.2E-01
	cg23838005	chr17	74868604	MGAT5B	-0.10	3.0E-06	8.2E-01
	cg00576027	chr17	11500910	DNAH9	-0.12	4.4E-06	8.9E-01
	cg01032978	chr1	20484113	RP3-340N1.2	-0.13	5.9E-06	8.9E-01
Cocaine only	cg26167583	chr11	65625524	CFL1	0.35	6.8E-08	4.1E-02
	cg19159404	chr17	79041535	BAIAP2	-0.35	9.9E-08	4.1E-02
	cg14976500	chr16	1210484	RP11-616M22.2	0.20	2.3E-07	5.8E-02
	cg21860679	chr12	89745673	DUSP6	0.21	3.0E-07	5.8E-02
	cg14527097	chr15	67329786	RP11-798K3.2	-0.32	3.5E-07	5.8E-02
Both drugs	cg18624512	chr6	151248413		0.15	4.6E-08	3.7E-02
	cg02596917	chr2	81423188		-0.09	8.2E-07	3.4E-01
	cg04490516	chr16	2255169	MLST8	0.08	2.5E-06	5.3E-01
	cg03546163	chr6	35654363	FKBP5	-0.16	2.7E-06	5.3E-01
	cg09182455	chr12	109116994	CORO1C	0.07	3.2E-06	5.3E-01

3.3.3 Joint analysis

We chose a correlation motif model with four groups of CpGs since this model has the lowest BIC among models with $K = 2, ..., 7$ (Figure 3.3). The correlation structure between groups of CpG sites and different phenotypes are shown in Figure 3.4.

In Figure 3.4, 905 CpG sites in Group 1 were found to be associated with any drug and heroin injection, and moderately associated with cocaine injection. 666 CpG sites in Group 2 were found to be associated with any drug and use of both drugs, and moderately associated with cocaine injection (enriched in inflammasome complex). 32,541 CpG sites in Group 3 were found to be associated with any drug and use of both drugs, and moderately associated with heroin injection. The rest 767,943 CpG

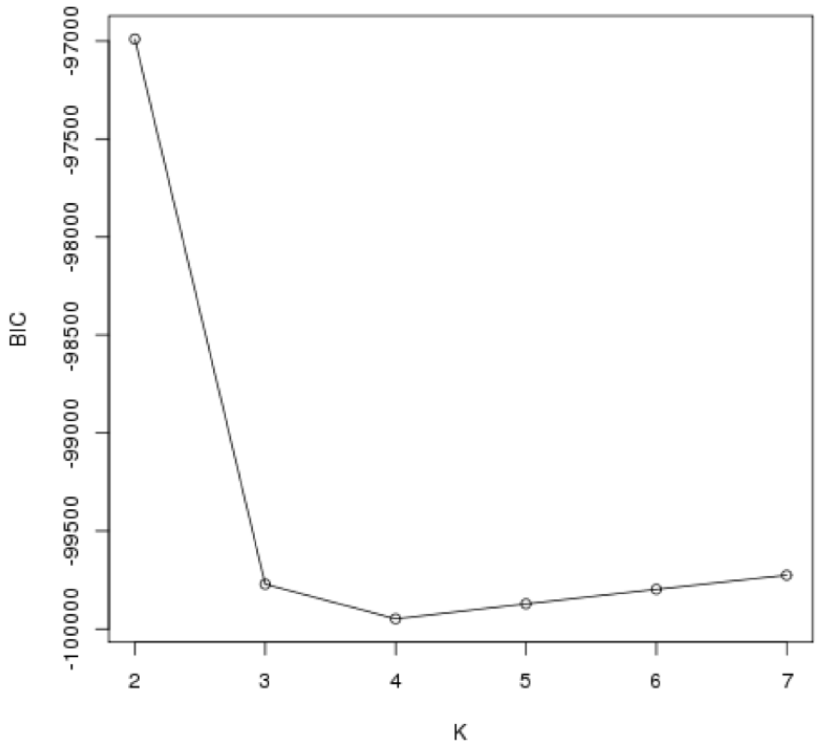


Figure 3.3: Number of group of CpGs in correlation motif model and BIC; select $K = 4$ as the final model with the lowest BIC

sites were not associated with any injection drug use.

Correlation structure in Joint Analyses

	Number of CpGs	Probability of each motif being significant			
		Any drugs	Heroin only	Cocaine only	Both drugs
Group 1	1,905	0.9995	0.9997	0.4301	0.0025
Group 2	666	0.9998	0.0025	0.3572	0.9997
Group 3	32,541	0.9999	0.3042	0.0007	0.9999
Group 4	767,943	3.0E-06	8.0E-06	3.3E-05	3.0E-06

Figure 3.4: Correlation motif structure based on $K = 4$

We compared the FDR between separate EWAS analyses and joint analyses after

borrowing information across different types of injection drug use in Table 3.3. For any drug injection, heroin only injection and co-use of heroin and cocaine injection, the top hits between separate and joint analyses remain the same. We acknowledge that the correlation motif method is developed under conditional independence assumption, and there is case overlap between any injection drug use and heroin only, cocaine only and co-use of heroin and cocaine. Thus, the results from the joint model should be interpreted carefully.

Table 3.3: Comparison of FDR on top hits between separate and joint analyses

Table: Comparison of Top Hits between Separate and Joint Analyses

Phenotype	Separate analyses		Joint analyses	
	Probe	FDR	Probe	FDR
Any drug	cg03012169	1.1E-01	cg03012169	3.0E-05
	cg03546163	2.4E-01	cg03546163	4.5E-05
	cg03067296	2.4E-01	cg03067296	6.9E-05
	cg10636246	4.2E-01	cg27407935	1.0E-04
	cg10564101	4.2E-01	cg10636246	1.1E-04
Heroin only	cg15989617	8.2E-01	cg15989617	1.1E-02
	cg22209773	8.2E-01	cg23838005	1.3E-02
	cg23838005	8.2E-01	cg18491269	1.5E-02
	cg00576027	8.9E-01	cg19481337	1.6E-02
	cg01032978	8.9E-01	cg22209773	1.6E-02
Cocaine only	cg26167583	4.1E-02	cg19481337	2.4E-02
	cg19159404	4.1E-02	cg06410104	3.0E-02
	cg14976500	5.8E-02	cg16956943	3.4E-02
	cg21860679	5.8E-02	cg15078654	3.6E-02
	cg14527097	5.8E-02	cg19159404	4.4E-02
Both heroin and cocaine	cg18624512	3.7E-02	cg18624512	1.6E-04
	cg02596917	3.4E-01	cg03546163	1.4E-03
	cg04490516	5.3E-01	cg09182455	1.5E-03
	cg03546163	5.3E-01	cg02596917	1.9E-03
	cg09182455	5.3E-01	cg12422199	2.0E-03

3.4 Discussion

There is a general lack of knowledge on how injection drug use affects DNA methylation in humans. Evidence on the effect of heroin and cocaine use in human and mice models has focused on either methylation on specific genes like OPRM1, or global methylation or demethylation.[117,118] The development of methylation profiling microarray technology allowed us to measure millions of CpG sites at lower cost and conduct epigenome-wide association studies(EWAS). EWAS in smoking and alcohol have revealed the utility of epigenetic sites as biomarker for smoking, maternal smoking and alcohol consumption.[21,119–121] It is highly possible that there exists epigenetic markers associated with injection drug use such as heroin and cocaine.

The separate joint analyses provided several CpG sites and genomic region of interest, and can be further explored by biological studies. Joint analyses also revealed groups of CpG sites that contribute to any injection drug use. However, compared to a recent EWAS analysis, there is no overlap between the identified CpG sites and genes.[122] Since the authors made comparison between lifetime injection drug use (and HCV+) vs. none injection drug use (and HCV-) among HIV infected people, which is different from our phenotype of interest, there might be underlying difference on the overall human conditions. With more data on DNA methylation on humans, meta-analysis can be used to aggregate across studies in this field and yield more reliable results.

3.5 Supplementary material

3.5.1 Supplementary figures and tables

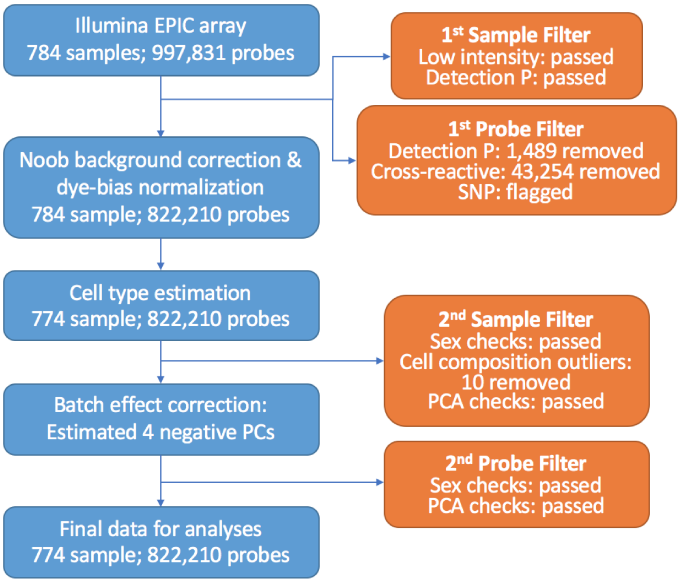


Figure 3.5: Overview of preprocessing pipeline

3.5.2 Supplementary methods

In many genetics and epigenetics studies, multiple phenotypic indicators or measures were collected for the sample. These phenotypes are usually correlated, for example, different types of drug use status for the past six months. Typically, we run separate genetics and epigenetics analyses on these phenotypes, ignoring the the potential correlation between the analyses. A joint analysis over these correlated phenotypes can be a better approach. By borrowing information across phenotypes,

we may detect genetic or epigenetic variants that cannot be discovered by separate analyses alone. By using the summary level statistics, we want to examine if there are groups of genetic or epigenetic variants of same patterns of association with the phenotypes. By accounting for the correlation across phenotypes, we can generate new statistics that may lead to new discovery and insight to the phenotype of interest.

Correlation motif in expression data

Developed by Dr. Hongkai Ji's group, the joint analysis method called correlation motif that can meta-analyze gene expression data in different studies.[114] An overview of the method is shown in Figure 3.6.

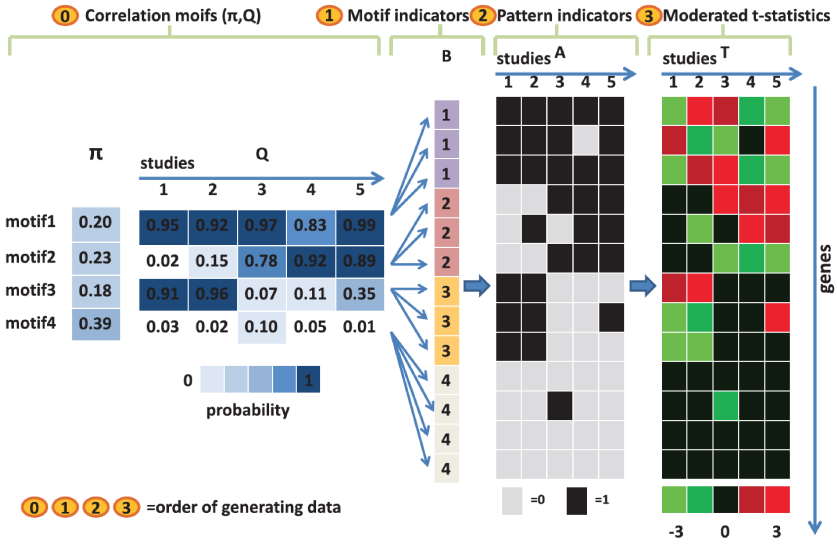


Figure 3.6: Illustration of correlation motif method

Let g indicates the number of genes and d indicates the number of studies in separate analyses. $\mathbf{T} = (t_{gd})$ is a $G \times D$ input matrix of summary test statistics from separate analysis. a_{gd} is the indicator on whether the gene a is deferentially

expressed in study d. The proposed method first finds K groups of genes that have similar differential expression pattern across different studies with the group label as b_g with probability $\pi = (\pi_1, ..., \pi_K)$, where $\pi_k = Pr(b_g = k)$ and $\sum_k \pi_k = 1$. The probability of any gene in class $b_g = k$ to be associated with study d is defined as $q_{kd} = Pr(a_{gd} = 1|b_g = k)$, which is the correlation structure between groups of genes and study. f_0 is the null distribution of t_{gd} and f_1 is the alternative distribution.

This method extracts correlation structure π_k, q_{kd} referred as correlation motif, and then calculates the posterior probability based on the correlation motif as prior and data from separate analyses. The joint probability is defined as follows:

$$Pr(\pi, \mathbf{Q}, \mathbf{A}, \mathbf{B}|\mathbf{T}) \propto \prod_{g=1}^{G}\prod_{k=1}^{K}\left\{\pi_k \prod_{d=1}^{D}\left[q_{kd} f_{d1}(t_{gd})\right]^{a_{gd}}\left[(1-q_{kd}) f_{d0}(t_{gd})\right]^{1-a_{gd}}\right\}^{\delta(b_g=k)} * \prod_{k=1}^{K}\pi_k \prod_{k=1}^{K}\prod_{d=1}^{D} q_{kd}(1-q_{kd})$$

where we use a Dirichlet prior for π and a beta prior $B(2, 2)$ for q_{kd}.

To determine what is the optimal number of group K, we use bayesian information criterion(BIC) as the criterion. The output is whether each gene g being significant in study d, i.e., a_{gd}, and is by default determined by the posterior probability greater than 0.5.

Correlation motif applied in DNA methylation data

To make correlation motif method applicable to different phenotypes in DNA methylation data, we make several changes to the model. First, we used p-values

instead of test statistics as summary statistic input from separate analyses. Second, the null distribution f_0 is assumed to be uniform (0,1), and the alternative distribution f_1 follows a beta distribution estimated by permutation. Third, we calculate the local FDR using 1-posterior probability to ensure we can make comparison between separate analyses and joint analyses. Details on how to formulate the model and run the EM algorithm is shown in the following sections.

3.5.2.1 E-step

The joint probability with priors are:

$$\begin{aligned} Pr(\pi, \mathbf{Q}, \mathbf{A}, \mathbf{B}|\mathbf{T}) \quad \propto \quad & \prod_{g=1}^{G}\prod_{k=1}^{K}\left\{\pi_k \prod_{d=1}^{D}\left[q_{kd} f_{d1}(t_{gd})\right]^{a_{gd}}\left[(1-q_{kd}) f_{d0}(t_{gd})\right]^{1-a_{gd}}\right\}^{\delta(b_g=k)} \\ & * \prod_{k=1}^{K}\pi_k \prod_{k=1}^{K}\prod_{d=1}^{D} q_{kd}(1-q_{kd}) \end{aligned}$$

The log-likelihood is:

$$\begin{aligned} \ln Pr(\pi, \mathbf{Q}, \mathbf{A}, \mathbf{B}|\mathbf{T}) \quad = \quad & \sum_{g=1}^{G}\sum_{k=1}^{K}\delta(b_g=k)\ln\pi_k \\ & + \sum_{g=1}^{G}\sum_{k=1}^{K}\delta(b_g=k)\{\sum_{d=1}^{D} a_g d[\ln q_{kd} + \ln f_{d1}(t_{gd})] \\ & + \sum_{d=1}^{D}(1-a_g d)[\ln(1-q_{kd}) + \ln f_{d0}(t_{gd})]\} \\ & + \sum_{k=1}^{K}\ln\pi_k + \sum_{k=1}^{K}\sum_{d=1}^{D}[\ln q_{kd} + \ln(1-q_{kd})] + constant \end{aligned}$$

Thus, the expectation of loglikehood with respect to $A, B|T, \hat{\pi}^{old}\hat{Q}^{old}$ is:

$$\begin{aligned}
Q(\pi, \mathbf{Q}|\hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}}) &= E_{old}\left[\ln Pr(\pi, \mathbf{Q}, \mathbf{A}, \mathbf{B}|\mathbf{T})\right] \\
&= \sum_{g=1}^{G}\sum_{k=1}^{K} \ln \pi_k E_{old}[\delta(b_g = k)] \\
&\quad + \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{d=1}^{D}[\ln q_{kd} + \ln f_{d1}(t_{gd})]E_{old}[\delta(b_g = k)a_{gd}] \\
&\quad + \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{d=1}^{D}[\ln(1 - q_{kd}) + \ln f_{d0}(t_{gd})]E_{old}[\delta(b_g = k)(1 - a_{gd})] \\
&\quad + \sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K}\sum_{d=1}^{D}[\ln q_{kd} + \ln(1 - q_{kd})] + constant
\end{aligned}$$

3.5.2.2 M-step

To obtain the maximum values after the E-step, we take the first order of partial derivative of $Q(\pi, \mathbf{Q}|\hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}})$ as follows:

$$\begin{aligned}
\frac{\partial Q\left(\pi, \mathbf{Q}|\hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}}\right)}{\partial \pi_k} &= 0 \\
\frac{\partial Q\left(\pi, \mathbf{Q}|\hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}}\right)}{\partial q_{kd}} &= 0
\end{aligned}$$

We have:

$$\begin{aligned}
\hat{\pi}_k^{new} &= \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1}{G + K} \\
\hat{q}_{kd}^{new} &= \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + 1}{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 2}
\end{aligned}$$

Details on how to derive the maximum is shown below:

To obtain $\hat{\pi}_k^{new}$, since $\sum_{k=1}^{K} \pi_k = 1$ and $E_{old}\left[\delta(b_g = k)\right] = Pr_{old}(b_g = k)$, we have:

$$\begin{aligned}
\frac{\partial Q\left(\pi, \mathbf{Q} | \hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}}\right)}{\partial \pi_k} &= 0 \\
\frac{1}{\pi_k} - \frac{1}{\pi_K} + \frac{1}{\pi_k} \sum_{g=1}^{G} Pr_{old}(b_g = k) - \frac{1}{\pi_K} \sum_{g=1}^{G} Pr_{old}(b_g = K) &= 0 \\
\frac{1}{\pi_k} \left[\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1\right] &= \frac{1}{\pi_K} \left[\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1\right] \\
\pi_k &= \frac{\pi_K \left[\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1\right]}{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1}
\end{aligned}$$

Since $\sum_{k=1}^{K} \pi_k = 1$, then:

$$\begin{aligned}
\sum_{k=1}^{K} \pi_k = \frac{\sum_{k=1}^{K} \pi_K \left[\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1\right]}{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1} &= 1 \\
\frac{\pi_K \left[\sum_{g=1}^{G} \sum_{k=1}^{K} Pr_{old}(b_g = k) + \sum_{k=1}^{K} 1\right]}{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1} &= 1 \\
\frac{\pi_K \left[\sum_{g=1}^{G} 1 + K\right]}{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1} &= 1 \\
\frac{\pi_K (G + K)}{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1} &= 1 \\
\pi_K &= \frac{\sum_{g=1}^{G} Pr_{old}(b_g = K) + 1}{G + K}
\end{aligned}$$

Thus,

$$\hat{\pi}_k^{new} = \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1}{G + K}$$

To obtain $\hat{q}_{kd}^{new}$, we have

$$\begin{aligned}
\frac{\partial Q\left(\pi, \mathbf{Q} | \hat{\pi}^{\mathbf{old}}, \hat{\mathbf{Q}}^{\mathbf{old}}\right)}{\partial \pi_k} &= 0 \\
\frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + 1}{q_{kd}} - \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 0) + 1}{1 - q_{kd}} &= 0
\end{aligned}$$

Thus,

$$\begin{aligned}\hat{q}_{kd}^{new} &= \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + 1}{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + \sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 0) + 2} \\ &= \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + 1}{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 2}\end{aligned}$$

By definition of $Pr_{old}(b_g = k)$, it is the total probability of genes assigned to motif k.

Thus,

$$\begin{aligned}Pr_{old}(b_g = k) &= \frac{\hat{\pi}_k^{(old)} \prod_{d=1}^{D} [\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1 - \hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})]}{\sum_{l=1}^{K} \hat{\pi}_l^{(old)} \prod_{d=1}^{D} [\hat{q}_{ld}^{(old)} f_{d1}(t_{gd}) + (1 - \hat{q}_{ld}^{(old)}) f_{d0}(t_{gd})]} \\ Pr_{old}(b_g = k, a_{gd} = 1) &= Pr_{old}(a_{gd} = 1 | b_g = k) * Pr_{old}(b_g = k) \\ &= \frac{\hat{q}_{kd}^{(old)} f_{d1}(t_{gd})}{\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1 - \hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})} * Pr_{old}(b_g = k)\end{aligned}$$

The posterior distribution is given by:

$$\begin{aligned}E(a_{gd} | \mathbf{T}, \hat{\pi}, \hat{\mathbf{Q}}) &= Pr(a_{gd} = 1 | \mathbf{T}, \hat{\pi}, \hat{\mathbf{Q}}) \\ &= \int_k Pr(b_g = k, a_{gd} = 1) \\ &= \sum_{k=1}^{K} \frac{\hat{q}_{kd}^{(old)} f_{d1}(t_{gd})}{\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1 - \hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})} * Pr_{old}(b_g = k)\end{aligned}$$

According to how we set up the problem, the log-likelihood is:

$$\begin{aligned}\ln Pr(\mathbf{T}|\pi,\mathbf{Q}) &= \sum_{g=1}^{G}\ln\left\{\sum_{k=1}^{K}\{\hat{\pi}_k^{(old)}\prod_{d=1}^{D}[\hat{q}_{kd}^{(old)}f_{d1}(t_{gd})+(1-\hat{q}_{kd}^{(old)})f_{d0}(t_{gd})]\}\right\}\\ &\quad+\sum_{k=1}^{K}\ln\pi_k+\sum_{k=1}^{K}\sum_{d=1}^{D}[\ln q_{kd}+\ln(1-q_{kd})]\end{aligned}$$

The BIC for K is:

$$\begin{aligned}BIC(K) &= -2\sum_{g=1}^{G}\ln Pr(\mathbf{T}|\pi,\mathbf{Q})+(K-1+K*D)*\ln G\\ &= -2\ln\left\{\sum_{k=1}^{K}\{\hat{\pi}_k^{(old)}\prod_{d=1}^{D}[\hat{q}_{kd}^{(old)}f_{d1}(t_{gd})+(1-\hat{q}_{kd}^{(old)})f_{d0}(t_{gd})]\}\right\}\\ &\quad+(K-1+K*D)*\ln G\end{aligned}$$

For missing data in phenotype d and CpG g, we have $f_{d1}(t_{gd})=1$ and $f_{d0}(t_{gd})=1$.

In $Pr_{old}(b_g=k)$, the numerator becomes:

$$\hat{\pi}_k^{(old)}[\hat{q}_{kd}^{(old)}f_{d1}(t_{gd})+(1-\hat{q}_{kd}^{(old)})f_{d0}(t_{gd})]=\hat{\pi}_k^{(old)}[\hat{q}_{kd}^{(old)}+(1-\hat{q}_{kd}^{(old)})]=\hat{\pi}_k^{(old)}$$

In $Pr_{old}(b_g=k,a_{gd}=1)$:

$$\begin{aligned}Pr_{old}(b_g=k,a_{gd}=1) &= \frac{\hat{q}_{kd}^{(old)}f_{d1}(t_{gd})}{\hat{q}_{kd}^{(old)}f_{d1}(t_{gd})+(1-\hat{q}_{kd}^{(old)})f_{d0}(t_{gd})}*Pr_{old}(b_g=k)\\ &= \hat{q}_{kd}*Pr_{old}(b_g=k)\end{aligned}$$

Thus, we can substitute $f_{d0}(t_{gd}), f_{d1}(t_{gd})$ with 1 for the missing observations, and it does not affect the final estimation. We also use the following formula to make sure a and b are close to zero, the estimation process will remain accurate.

$$\begin{aligned}\log(a+b) &= \log(\max(a,b)\left(\frac{a}{\max}+\frac{b}{\max}\right) = \log\max + \log\left(\frac{a}{\max}+\frac{b}{\max}\right)\\ &= \log\max + \log\left(e^{\log a-\log\max}+e^{\log b-\log\max}\right)\end{aligned}$$

3.5.2.3 Implementation of E-M algorithm

1. Choose initial values for $\pi_k^{(0)}$ and $q_{kd}^{(0)}$.

2. Calculate $Q^{(1)}(\pi, \mathbf{Q}|\hat{\pi}^{(\mathbf{0})}, \hat{\mathbf{Q}}^{(\mathbf{0})})$ based on $\pi_k^{(0)}$ and $q_{kd}^{(0)}$:

(1) $a_{1,kdg} = \ln[\hat{q}_{kd}^{(old)} f_{d1}(t_{gd})]$; for missing data, $a_{1,kdg} = \ln \hat{q}_{kd}^{(old)}$ (3d-array)

$a_{2,kdg} = \ln[(1-\hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})]$; for missing data, $a_{2,kdg} = \ln(1-\hat{q}_{kd}^{(old)})$ (3d-array)

(2) $a_{max,kdg} = pmax(a_{1,kdg}, a_{2,kdg})$ (3d-array)

$e_{1,kdg} = e^{a_{1,kdg}-a_{max,kdg}}$ (3d-array)

$e_{2,kdg} = e^{a_{2,kdg}-a_{max,kdg}}$ (3d-array)

(3) $logp_{kg} = \ln \hat{\pi}_k^{(old)} + \sum_{d=1}^{D}\{\ln[\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1-\hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})]\}$

$= \ln \hat{\pi}_k^{(old)} + \sum_{d=1}^{D}\{\ln[e^{a_{1,kdg}} + e^{a_{2,kdg}}]\}$

$= \ln \hat{\pi}_k^{(old)} + \sum_{d=1}^{D}\{\ln[e^{a_{1,kdg}-a_{max,kdg}} + e^{a_{2,kdg}-a_{max,kdg}}] + a_{max,kdg}\}$

$= \ln \hat{\pi}_k^{(old)} + \sum_{d=1}^{D}[\ln(e_{1,kdg}+e_{2,kdg}) + a_{max,kdg}]$ (2d-array)

$logpmax_{kg} = pmax(logp_{kg})$ (2d-array)

(4) $Pr_{old}(b_g = k) = \frac{e^{logp_{kg} - logpmax_{kg}}}{\sum_{l=1}^{K} e^{logp_{lg} - logpmax_{lg}}}$ (probability of a specific g belongs to motif k, output, 2d-array)

(5) $Pr_{old}(b_g = k, a_{gd} = 1) = \frac{e_{1,kdg}}{e_{1,kd} + e_{2,kdg}} * Pr_{old}(b_g = k)$; (3d-array)

$Pr_{old}(b_g = k, a_{gd} = 0) = \frac{e_{2,kdg}}{e_{1,kd} + e_{2,kdg}} * Pr_{old}(b_g = k)$; (3d-array)

(6) $Q^{(1)}(\pi, \mathbf{Q}|\hat{\pi}^{(\mathbf{0})}, \hat{\mathbf{Q}}^{(\mathbf{0})}) =$

$\sum_{k=1}^{K} \ln \pi_k^{(0)} \left(\sum_{g=1}^{G} Pr_{old}^{(0)}(b_g = k) + 1 \right)$

$+ \sum_{g=1}^{G} \sum_{k=1}^{K} \sum_{d=1}^{D} [a_{1,kdg} Pr_{old}^{(0)}(b_g = k, a_{gd} = 1) + a_{2,kdg} Pr_{old}^{(0)}(b_g = k, a_{gd} = 0)]$

$+ \sum_{k=1}^{K} \sum_{d=1}^{D} [\ln q_{kd}^{(0)} + \ln(1 - q_{kd}^{(0)})] + constant$

3. Calculate $\pi_k^{(1)}$ and $q_{kd}^{(1)}$ based on $Q^{(1)}(\pi, \mathbf{Q}|\hat{\pi}^{(\mathbf{0})}, \hat{\mathbf{Q}}^{(\mathbf{0})})$

$\pi_k^{(1)} = \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 1}{G + K}$ (1d-array)

$q_{kd}^{(1)} = \frac{\sum_{g=1}^{G} Pr_{old}(b_g = k, a_{gd} = 1) + 1}{\sum_{g=1}^{G} Pr_{old}(b_g = k) + 2}$ (2d-array)

4. Repeat 2-3 for a maximum of n times, or when none of the parameters in π and $\mathbf{Q}$ changes by more than 0.1%.

5. The posterior distribution is: $E(a_{gd}|\mathbf{T}, \hat{\pi}, \hat{\mathbf{Q}}) = \sum_{k=1}^{K} \frac{e_{1,kdg}}{e_{1,kd} + e_{2,kdg}} * Pr_{old}(b_g = k)$

6. The marginal log-likelihood is:

$$\begin{aligned}
\ln Pr(\mathbf{T}|\pi, \mathbf{Q}) &= \sum_{g=1}^{G} \ln \left\{ \sum_{k=1}^{K} \{ \hat{\pi}_k^{(old)} \prod_{d=1}^{D} [\hat{q}_{kd}^{(old)} f_{d1}(t_{gd}) + (1 - \hat{q}_{kd}^{(old)}) f_{d0}(t_{gd})] \} \right\} \\
&\quad + \sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K} \sum_{d=1}^{D} [\ln q_{kd} + \ln(1 - q_{kd})] \\
&= \sum_{g=1}^{G} \ln \left\{ \sum_{k=1}^{K} e^{logp_{kg}} \right\} + \sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K} \sum_{d=1}^{D} [\ln q_{kd} + \ln(1 - q_{kd})] \\
&= \sum_{g=1}^{G} \left\{ logpmax_{kg} + \ln \sum_{k=1}^{K} e^{logp_{kg} - logpmax_{kg}} \right\} + \sum_{k=1}^{K} \ln \pi_k \\
&\quad + \sum_{k=1}^{K} \sum_{d=1}^{D} [\ln q_{kd} + \ln(1 - q_{kd})]
\end{aligned}$$

7. The BIC for K is:

$$BIC(K) = -2 \sum_{g=1}^{G} \ln Pr(\mathbf{T}|\pi, \mathbf{Q}) + (K - 1 + K * D) * \ln G$$

3.5.2.4 EM algorithm for estimating f_0, f_1

Set up:

1. Observed empirical p-values from permutation $\mathbf{X}$.

2. Missing probability of belonging to null distribution for each CpG site $\mathbf{Z}$.

3. Parameters to be estimated as listed below:

$f = \pi_0 f_0 + \pi_1 f_1$

(1) f is the overall distribution.

(2) π_0 is the probability of being in the null distribution (expected to be > 0.99); needs to be estimated.

(3) π_1 is the probability of being in the alternative distribution (expected to be < 0.01); needs to be estimated.

(4) f_0 is the null distribution with Unif(0,1)

(5) f_1 is the alternative distribution with Beta(α,β) ; α, β needs to be estimated.

Steps:

1. Choose initial values for $\pi_0, \pi_1, \alpha, \beta$.

$\pi_0^{(0)} = 0.99, \pi_1^{(0)} = 0.01, \alpha^{(0)} = 1, \beta^{(0)}$=30

2. Compute f based on initial values.

3. Compute posterior probability for each CpG site belonging to f_0 as $\mathbf{Z}$.

Denote the model based p-value for CpG_i is p_i:

$$Z_i = \frac{\pi_0^{(0)} f_0(p_i)}{\pi_0^{(0)} f_0(p_i) + \pi_1^{(0)} f_1(p_i)}$$

4. Use $\mathbf{Z}$ to estimate $\pi_0^{(1)}, \pi_1^{(1)}, \alpha^{(1)}, \beta^{(1)}$.

$$\pi_0^{(1)} = \frac{\sum_i Z_i}{n},\ \pi_1^{(1)} = 1 - \frac{\sum_i Z_i}{n}$$

$$\alpha^{(1)} = \frac{\mu^2(1-\mu)}{\sigma^2} - \mu,$$

$$\beta^{(1)} = \frac{\mu(1-\mu)^2}{\sigma^2} + \mu - 1,$$

$$w_i = (1 - Z_i) / \sum_j (1 - Z_j)$$

$$\hat{\mu} = \sum_i w_i p_i$$

$$\hat{\sigma}^2 = \sum_i w_i (p_i - \hat{\mu})^2$$

5. Repeat 3-4 until converge.

Chapter 4

Epigenetics and HIV

4.1 Introduction

The incidence of HIV has declined between 2011 and 2015.[123] With appropriate antiretroviral treatment and care, the life expectancy of HIV-infected patients even approaches uninfected individuals.[124] HIV has become a chronic disease, but latent HIV induces chronic inflammation and has led to HIV many related commorbidity and complications such as cardiovascular disease and dementia.[125] Chronic HIV infection has a long-term impact on the immune system, but the exact biological mechanisms affected are not known.

Epigenetics studies open a door to better understanding the changes in epigenetic profiles with chronic HIV infection. Acute HIV infection on CD4+ and CD8+ cell lines is proved to be associated with global methylation changes in vitro.[126] There is

also evidence that HIV latency is regulated by DNA methylation and histone modification.[46,127,128] DNA methylation level in viral promoter such as the HIV long terminal repeat(LTR) region affect viral gene expression.[128] Thus, epigenetic analyses may provide potential drug targets to remove latent virus reservoir.[129]

Recent studies have shown that there are epigenetic signals associated with HIV infection in the NLRC5 and HLA genes.[43,44,130] The top CpG signal associated gene region in these studies has also presented in HIV integration sites.[45] It will be interesting to further study the epigenetics signal with better coverage of CpG sites and with careful adjustment for cell type heterogeneity between HIV positive and negative samples.

Studies have shown that injection drug users have much higher risk of HIV infection[32] and worse HIV outcome during treatment initiation compared with non-IDUs. Notably, injection drug use may affect HIV progression,[33,34] and may lead to shortened disease-free survival time in HAART treatment.[34] Injection drug users were shown to be less likely to engage in HIV care.[131] Among frequently injected drugs, heroin is known to have immunosuppressive effects in human by affecting T lymphocyte functions and inhibiting T cell signaling.[35–37] It would be also interesting to examine how injection drug use affects HIV latency.

4.2 Method

4.2.1 Study samples

The ALIVE study is an on-going prospective cohort study characterizing the incidence and natural history of HIV infection among injection drug users (IDUs) in Baltimore, MD from 1988.[27] At each 6-month visit, clinical, behavioral and laboratory data such as HIV infection status and injection drug use were assessed for the participants. Blood was obtained from 288 current IDUs, resampled after cessation and then again after relapse (total samples = 774). HIV status did not change across visits.

In order to avoid repeated measurement on the same subjects with the same HIV status and confounding by injection drug use, we randomly select one visit without any past six month injection drug use, and the total sample size is 281 subjects with 182 HIV positives, 99 HIV negatives. For sensitivity analyses, we also randomly select one visit with any past six month injection drug use, and the sample size is 218 subjects with 147 HIV positives, 71 HIV negatives.

The study participants went through HIV serology screening at baseline. HIV serology status was assessed at each study visit for HIV negative participants, whereas CD4+, CD8+ count and HIV viral load testing were performed at each study visit for HIV positive participants. Past six injection drug use status and type, smoking patterns, and ART use information were obtained by computer-administered standard-

ized questionnaires. The study design had been described in other literature.[27,115,116]

4.2.2 DNA methylation measurement and preprocessing

Peripheral blood mononuclear cell DNA was isolated with the Qiagen DNeasy kit and bisulfite converted with Zymo EZ methylation gold kit at the Johns Hopkins University Center for Inherited Disease Research. Bisulfite treated DNA was run on the Illumina Infinium MethylationEPIC BeadChip.

The *minfi* package (Bioconductor) was used to process raw Illumina image files into noob preprocessed methylation beta values.[91] Cell composition on CD4+, CD8+, natural killer cells, monocytes, granulocytes and B cells were estimated based on the method described in Houseman et al.[96] and implemented in *minfi*.[91,103] Samples with low intensity, inconsistency on predicted and observed sex, and outliers in estimated cell composition were removed for quality control. Probes with low intensity or that are known to cross-hybridize were excluded. 774 samples and 822,210 probes were used in the final analyses. Batch effect was adjusted by top 4 PCs from negative control features that are only correlated with technical variations.[101] The pipeline of proprecessing is in supplementary figure 4.7.

Since HIV infection is known to be associated with lower CD4+ count and higher CD8+ count, in addition to the estimated cell composition, we also adjusted for top

6 PCs of 1,000 most significant CpGs that are positively associated with CD8+ but negatively associated with CD4+ by using the R package *FlowSorted.Blood.450k* on Illumina HumanMethylation data on sorted blood cell populations.[132]

4.2.3 Statistical analysis

M value is used in the analysis since the M-value distribution is closer to the assumption of normality and less variable. Researchers reported that the use of M value usually leads to better detection rates and true positive rates compared with the beta value.[93]

We used *limma* to run single site association analyses with HIV infection status across the epigenome for the 822,210 CpG sites, adjusting for gender, race, age, smoking status, cell composition, PCs for batch effect, and additional 6PCs representing more CD8+ and less CD4+. The model is stated below:

$$
\begin{aligned}
M_i \;=\; & \beta_0 + \beta_1 \text{HIV}_i + \beta_2 \text{gender}_i + \beta_3 \text{race}_i + \beta_4 \text{age}_i + \beta_5 \text{smoking}_i + \beta_6 \text{CD4}_i + \beta_7 \text{CD8}_i \\
& + \beta_8 \text{natural killer cells}_i + \beta_9 \text{monocytes}_i + \beta_{10} \text{granulocytes}_i + \beta_{11} \text{B cells}_i \\
& + \beta_{12} \text{negative control PC1}_i + \beta_{13} \text{negative control PC2}_i + \\
& \beta_{14} \text{negative control PC3}_i + \beta_{15} \text{negative control PC4}_i \\
& + \beta_{16} \text{CD8/CD4 PC1}_i + \beta_{17} \text{CD8/CD4 PC2}_i + \beta_{18} \text{CD8/CD4 PC3}_i \\
& + \beta_{19} \text{CD8/CD4 PC4}_i + \beta_{20} \text{CD8/CD4 PC5}_i + \beta_{21} \text{CD8/CD4 PC6}_i
\end{aligned}
$$

For sensitivity analyses among 218 injection drug use visits, we additionally used type of injection drug use (heroin only injection, cocaine only injection and other) as covariates.

The gene ontology enrichment analysis is done using in GO(Gene Ontology) database[106] by the R pacakage *missMethyl*, with prior to correction for sampling bias.[108,109] We used the epigenetic clock to estimate the biological age by the method described in Horvath et al.[110]

4.3 Results

4.3.1 Cell composition

There were significant differences on CD4+ cell proportion and CD8+ cell proportions by HIV status (Figure 4.1). HIV infected majorly on CD4+ cells, and induced decrease in CD4+ cells and increase in CD8+ cells. These changes were reflected in estimated cell proportions in Figure 4.1.

We obtained measured CD4+ and CD8+ count and proportions for 230 HIV+ visits and 39 HIV- visits, and made comparison between estimated CD4+/CD8+ ratio vs. measured CD4+/CD8+ ratio in Figure 4.2. The estimated CD4+/CD8+ ratio is similar to the measured ratio, but there are many outliers and there is significant difference by HIV between estimated and measured CD4+/CD8+ ratio (p=0.047). This indicates that the cell proportion estimation procedure may underestimate CD4+ or

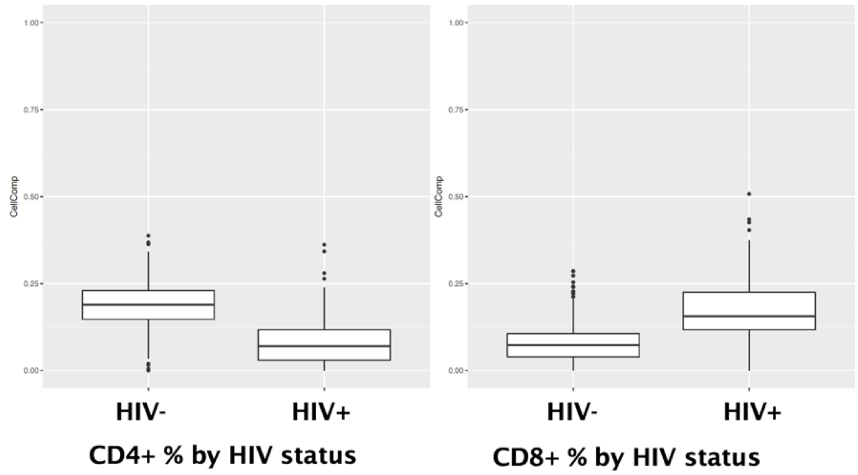


Figure 4.1: Estimated CD4+ and CD8+ proportion by HIV infection

overestimate CD8+ among HIV+.

4.3.2 EWAS on HIV among non-injection visits

For 281 subjects, we selected one "no injection" observation per subject to remove the effect of current injection on methylation signature, as illustrated in Figure 4.3. The sample characteristics is shown in Table 4.1. About 2/3 of the participants were male, and 35.2% of them were HIV positive. 92% of the samples were from African Americans. The average age of the sample visits was around 49.

The QQ-plot of HIV EWAS among no injection observations was shown in Figure 4.4. No apparent inflation was observed ($\lambda = 0.99$) after we accounted for PCs for CpGs that are negatively associated with CD4+ and positively associated with CD8+.

CpG sites associated with the NLRC5 and C12orf32 genes were ranked highest

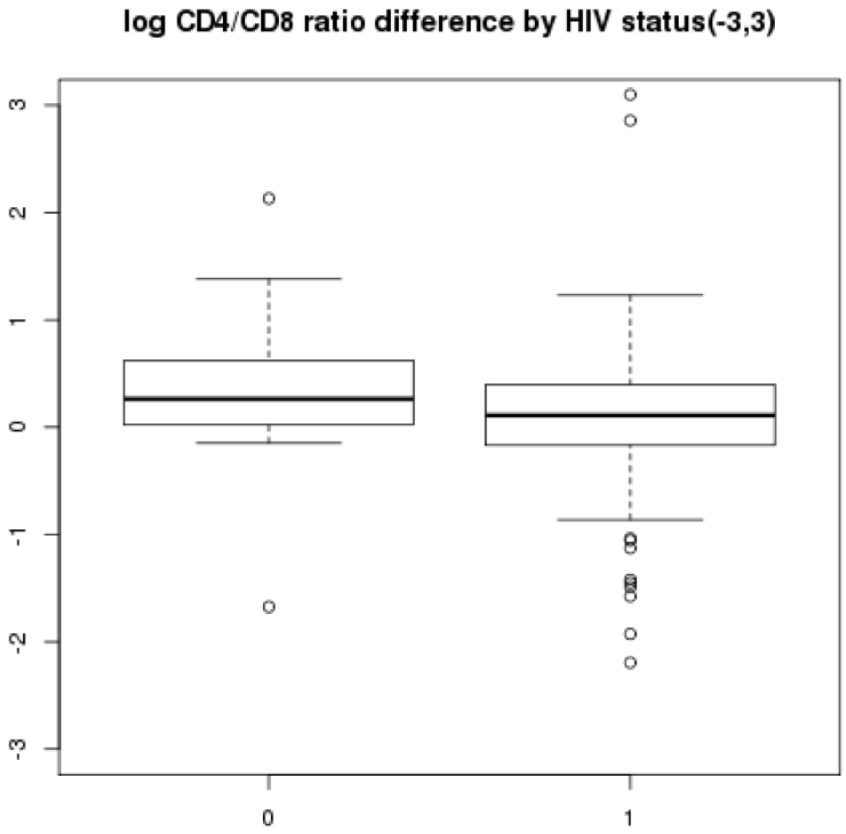


Figure 4.2: Difference between estimated and measured log CD4+/CD8+ ratio

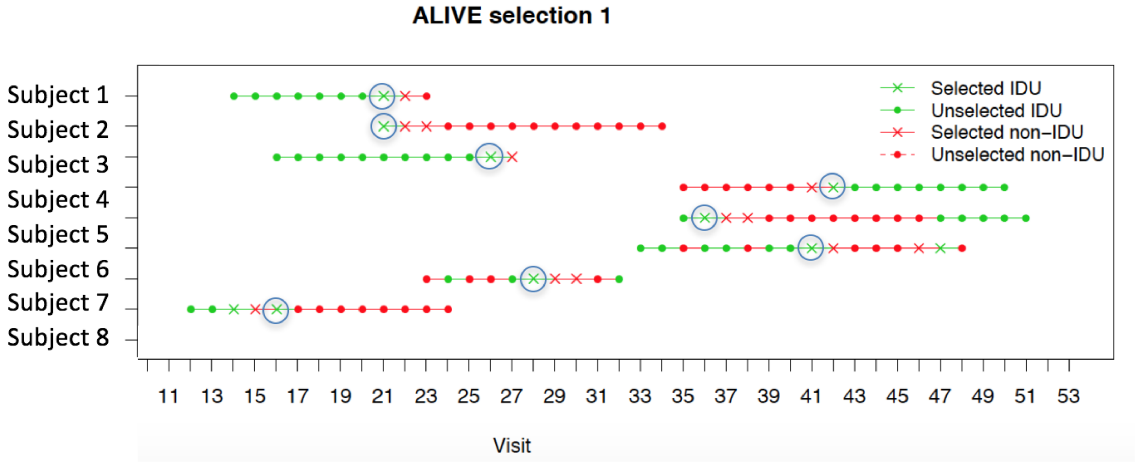


Figure 4.3: Illustration of study design; only subject's visits in red(cessation) were selected

Table 4.1: Demographics on non-injection visits

Table: Study sample characteristics (n=281)

	Mean(se)	N(%)
Average age	49.5±7.1	
Gender		
Male	193	68.7%
Female	88	31.3%
Race		
Caucasian/non-hispanic	15	5.3%
African Americans/non-hispanic	258	91.8%
Other	8	2.8%
HIV status		
Positive	99	35.2%
Negative	182	64.8%

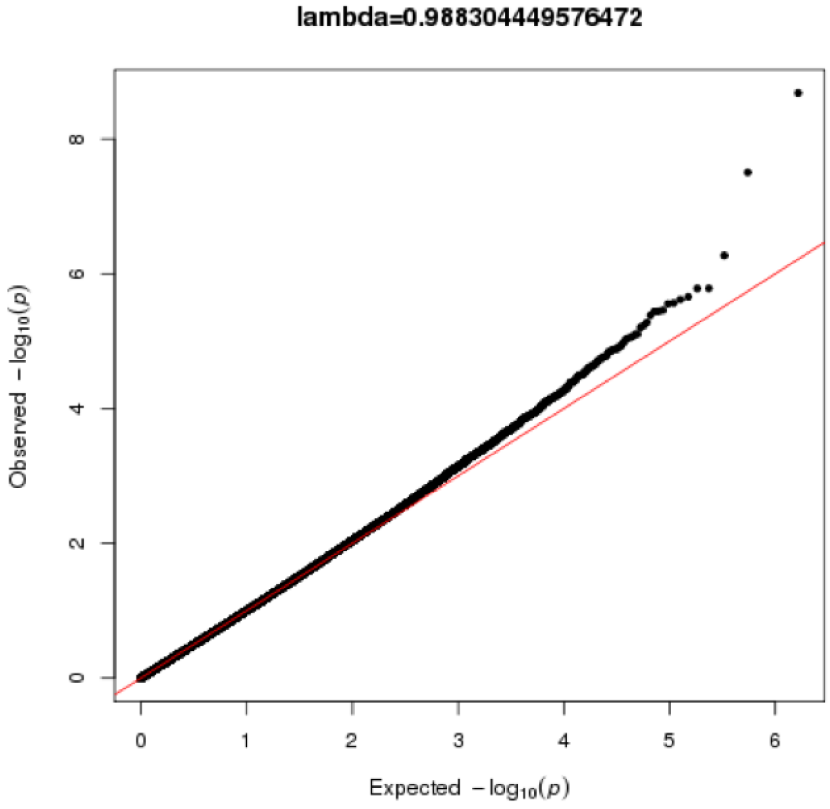


Figure 4.4: QQ plots of HIV among non-injection visits

for association with HIV status and remained significant after correction for multiple testing (Table 4.2). NLRC5 is involved in cytokine response and antiviral immunity, and the C12orf32 gene is associated with DNA double-strand break repair.

Table 4.2: Top hits in HIV among non-injection visits

Table: Top 10 CpG sites associated with HIV (n=281; no injection drug use)

Probe	CHR	BP	Gene	p-value	BH FDR
cg07839457	chr16	57023022	NLRC5	2.08E-09	0.002
cg12051710	chr12	2989070	C12orf32	3.14E-08	0.013
cg15331332	chr6	29692111	HLA-F	5.34E-07	0.146
cg23227370	chr7	45243040		1.65E-06	0.249
cg18544413	chr12	65019442	RASSF3	1.66E-06	0.249
cg16565442	chr2	97406324	LMAN2L	2.19E-06	0.249
cg10129948	chr11	57085849	TNKS1BP1	2.42E-06	0.249
cg10055090	chr14	1.04E+08	EIF5	2.72E-06	0.249
cg12941231	chr17	7296844	PLSCR3	2.80E-06	0.249
cg19097500	chr1	61542329	NFIA	3.47E-06	0.249

4.3.3 Sensitivity analysis: EWAS on HIV among injection visits

For 218 subjects, we selected one "any injection drug use" observation per subject. The sample characteristics is shown in Table 4.3. 19 subjects only injected cocaine, 64 subjects only injected heroin, while the rest of 135 subjects injected both heroin

and cocaine. About 70% of the participants were male, and 32.6% of them were HIV positive. 90.8% of the samples were from African Americans. The average age of the sample visits was around 48.

Table 4.3: Demographics on injection visits

Table: Study sample characteristics (n=218)		
	Mean(se)	**N(%)**
Average age	48.6±7.8	
Gender		
Male	153	70.2%
Female	65	29.8%
Race		
Caucasian/non-hispanic	13	6.0%
African Americans/non-hispanic	198	90.8%
Other	7	3.2%
HIV status		
Positive	71	32.6%
Negative	147	67.4%
Injection Drug type		
Heroin only	64	29.4%
Cocaine only	19	8.7%
Both drugs	135	61.9%

The QQ-plot of HIV EWAS among injection observations was shown in Figure 4.5. No apparent inflation was observed ($\lambda = 1.07$) after we account for PCs for CpGs that are negatively associated with CD4+ and positively associated with CD8+.

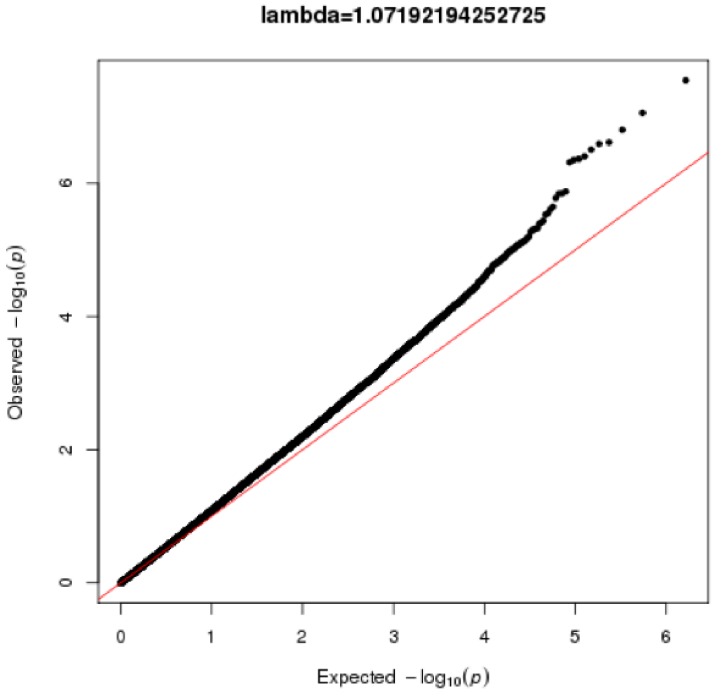


Figure 4.5: QQ plots of HIV among injection visits

CpG sites associated with the UQCRC2, TNF, and NLRC5 genes were ranked highest for association with IDU status and remained significant after correction for multiple testing (Table 4.4). NLRC5 is also found significant among no injection visits. By comparing the results from two analyses by CATplot in Figure 4.6, about 30% of the top 50 ranked HIV associated CpG sites between no injection and injection observations are the same.

Table 4.4: Top hits in HIV among injection visits

Table: Top 10 CpG sites associated with HIV (n=218; injection drug use)

Probe	CHR	BP	Gene	p-value	BH FDR
cg05925101	chr16	21991385	UQCRC2	2.85E-08	0.023
cg10717214	chr6	31543557	TNF	8.80E-08	0.036
cg07839457	chr16	57023022	NLRC5	1.57E-07	0.040
cg24208604	chr16	88976160	CBFA2T3	2.42E-07	0.040
cg01625368	chr11	33890924	LMO2	2.56E-07	0.040
cg23387863	chr15	77472416	SGK269	3.12E-07	0.040
cg23585168	chr6	32178215	NOTCH4	3.95E-07	0.040
cg14139935	chr9	1.3E+08	LRSAM1	4.31E-07	0.040
cg01962509	chr2	2.32E+08		4.45E-07	0.040
cg18387107	chr12	3980774	PARP11	4.83E-07	0.040

4.4 Discussion

The epigenome-wide association study on HIV from white blood cell samples are challenging because chronic HIV infection is closely associated with increased CD4+

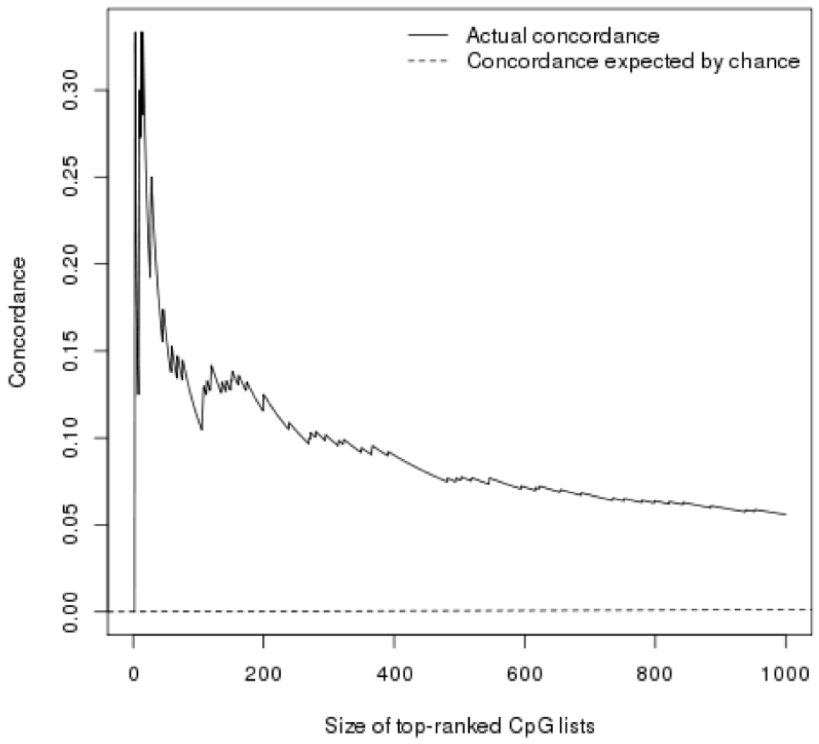


Figure 4.6: Concordance of top ranked CpG sites between HIV EWAS non-injection visits and injection visits

proportion and decreased CD8+ proportion. CpG sites associated with HIV may just be reflective of cell proportion. To avoid false positives from cell composition epigenetic markers, we penalized CpG sites that are positively associated with CD4+ cells and negatively associated with CD8+ cells.

The CpG site associated with NRLC5 gene come out as the top ranked gene among the HIV analyses in both non-injection and injection analyses. This gene has also been reported in other independent HIV EWAS studies.[43,44] Interestingly, the NRLC5 gene has been reported to be one of the persistent HIV integration sites on

chronic HIV infection patients.[45] This change in DNA methylation may be a result of either chronic infection by HIV virus, or HIV integration near this gene. Further biological experiments on this gene can shed light on how chronic HIV infection affects the immune system.

It is also interesting to explore how injection drug use modify chronic HIV infection's effect on the immune system. There is about 30% concordance between the non-injection visits and injection visits, indicating interaction effect of injection drug use and chronic HIV infection on DNA methylation in blood cells. Integrating injection drug use methylation markers in chronic HIV infection research may be interesting.

4.5 Supplementary figures and tables

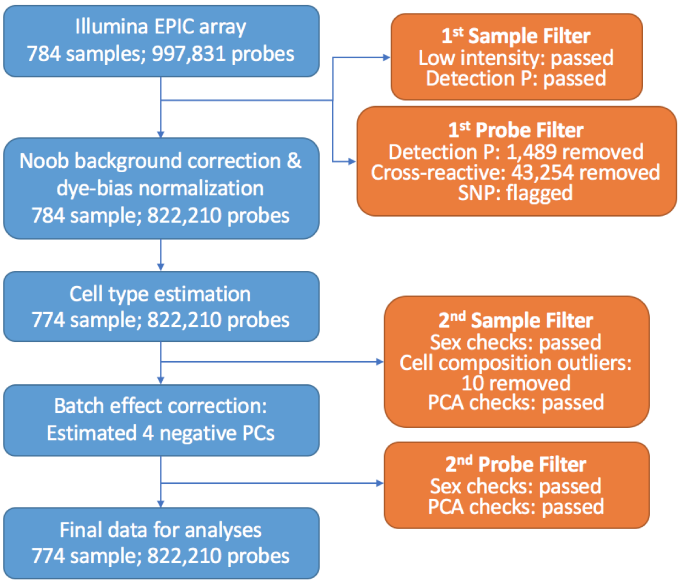


Figure 4.7: Overview of preprocessing pipeline

Chapter 5

Integrating Brain Expression Quantitative Loci in the Autism Spectrum Disorder Genome-wide Association Study

5.1 Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by significant social impairment and repetitive behaviour.[47] ASD is highly heritable, suggesting genetic factors play a major role in its etiology.[54] However, there is limited success in genome-wide association study (GWAS) on ASD due to small sample size.[12]

Many studies on psychiatric genetics, including ASD, studies have shown that many of the genes implicated in polygenic risk are involved in regulating gene expression and translation in the brain. These genetic variants are known as Brain expression quantitative trait loci, or Brain eQTLs. In the context of GWAS, we specifically refer to SNPs that are brain eQTLs as Brain eSNPs. These Brain eSNPs may pinpoint functionally important findings in GWAS and elucidate biological pathways associated with ASD.[67] Recent literature has shown that brain-informed eQTL annotation of genetic variants can enable us to discover novel variants in ASD.[133] However, due to limited sample sizes, studies on ASD GWAS often fail to report significant genetic variants. One approach to reducing the burden of multiple testing is to subset the scope of candidate SNPs in a GWAS panel to brain eSNPs. This in turn may help us detect functionally relevant SNPs. Further subsetting brain eSNPs to those specific to brain region may help us pinpoint brain regions of interest in ASD etiology. To our knowledge, no one has yet performed an ASD GWAS using candidate brain eSNPs informed by brain eQTL studies. We propose a brain eSNPs informed analysis on our GWAS sample, the Study to Explore Early Development.

5.2 Method

5.2.1 Overview

We describe the GWAS samples we are going to use, how we obtained brain eQTLs from existing literature, and how we created pruned all SNPs, brain eSNPs and brain region specific eSNPs subsets. Genome-wide association study (GWAS) Sample The Study to Explore Early Development (SEED). The SEED study consists of 1,309 samples, with 584 Autism Spectrum Disorder (ASD) cases and 725 normal controls with multiple ancestry, and over 30 million single-nucleotide polymorphisms (SNPs) that passed quality control (QC). The genotyping platform used was either the Illumina Omni1M Quad or the Affymetrix Kaiser Axiom array. The QC measures included removal of samples with low call rates ($< 98\%$), sex discrepancies, inappropriate relatedness ($\hat{\pi} > 0.2$), and/or excess heterozygosity or homozygosity, removal of SNPs with a minor allele frequency less than (MAF$< 5\%$), and flagging of SNPs statistically significant ($p < 10^{-6}$) for not being in Hardy Weinberg Equilibrium in ancestry-stratified control samples. Following application of QC filters, phasing was performed using SHAPEIT followed by imputation against the 1000 Genome Project panel using IMPUTE2, resulting in over 30 million SNPs per sample. Ten principal components representing the multiple genetic ancestry of each sample were used to adjust for ancestry in GWAS analyses.

5.2.2 Brain eSNP Data Sets

Brain eSNPs were extracted from 6 published brain eQTLs studies, and the brain tissues were from neurologically normal individuals and those with psychiatric disorders in Supplementary Table 5.2,[65,134–138] with annotation for 11 different brain tissue types- cerebellar cortex(CRBL), frontal cortex(FCTX), hippocampus(HIPP), inferior olivary nucleus(MEDU), occipital cortex(OCTX), pons(PONS), putamen(PUTM), substantianigra(SNIG), temporal cortex(TCTX), thalamus(THAL), intralobular white matter(WHMT).To get greater coverage of those Brain eSNPs in SEED GWAS, we obtained proxy SNPs($r^2 > 0.80$, 1000 genome as reference panel) for the eSNPs not found in SEED by using SNAP(http://archive.broadinstitute.org/mpg/snap/ldsearch.php). Together with the proxy SNPs, we obtained a Brain eSNPs subset of 41,556 eSNPs, as shown in the flow chart in Supplementary Figure 5.3.

5.2.3 Linkage disequilibrium(LD) pruning

We ran LD pruning by using the plink default method (variance inflation factor(VIF) with window size of 50, step size of 5 and VIF threshold at 2) for the brain eSNP subset and brain tissue specific eSNP subsets(Supplementary Figure 5.3). The final pruned brain eSNP subset contained 5,852 eSNPs in SEED. To make comparable all LD-pruned SNP subsets in SEED, we used Priority Pruner on all SNPs in SEED and with the pruned 5,852 eSNPs forced in Supplementary Figure 5.3. The

brain tissue specific eSNPs in SEED are shown in Table 5.1.

In order to assess whether brain eQTL informed subsets improves our ability to detect significant genetic variants, we compared the QQ plots between all pruned SNPs vs. pruned brain eSNPs, in SEED, we ran similar LD pruning for all SNPs with the 5,852 pruned eSNPs forced in by using Priority Pruner (http://prioritypruner.sourceforge.net/) with parameters equivalent to plink LD pruning (Supplementary Figure 5.3). A final count of pruned SNPs is shown in Table 5.1.

5.2.4 SEED GWAS subsetting to brain eSNPs

We used the pruned subsets from Table 5.1 and applied them in SEED GWAS analyses in plink (version 1.9) and R (version 3.3.1). QQ plots were examined for each SNP subsets in order to see whether there are significant SNPs detected in the subset. We calculated the adjusted significance level according to Bonferroni correction, and listed the top SNP with the lowest p-value in each subset. Annotation of the SNP and gene function is based on the NCBI databses dbSNP[139] and RefSeq.[140]

5.3 Results

5.3.1 Summary of Brain eSNPs subsets in SEED GWAS

A final summary of the number of SNPs in each brain eSNPs subset, based on SEED GWAS SNPs, is shown in Table 5.1.

Table 5.1: SEED GWAS Brain SNPs subsets overview

Table 1: SEED GWAS pruned SNPs overview

Category	Before pruning	After pruning (MAF>0.05, missingness<0.02)	Adjusted alpha according to pruned SNPs	Top SNP	
All SNPs	3827921	666,071	7.51E-08	rs57396002	1.12E-06
Brain eSNPs	41556	5852	8.54E-06	rs9876540	8.88E-05
CRBL	14395	1892	2.64E-05	rs1132306	5.23E-04
FCTX	10722	996	5.02E-05	rs12622456	1.97E-03
HIPP	1596	434	1.15E-04	rs10888841	5.03E-03
MEDU	1723	350	1.43E-04	rs75428054	3.68E-03
OCTX	1494	356	1.40E-04	rs11557043	3.88E-03
PONS	4094	327	1.53E-04	rs10736174	1.60E-02
PUTM	899	218	2.29E-04	rs2068946	3.58E-03
SNIG	667	193	2.59E-04	rs10888841	5.03E-03
TCTX	7107	886	5.64E-05	rs9876540	8.88E-05
THAL	1280	343	1.46E-04	rs116629141	4.69E-03
WHMT	2619	574	8.71E-05	rs7081640	1.81E-03

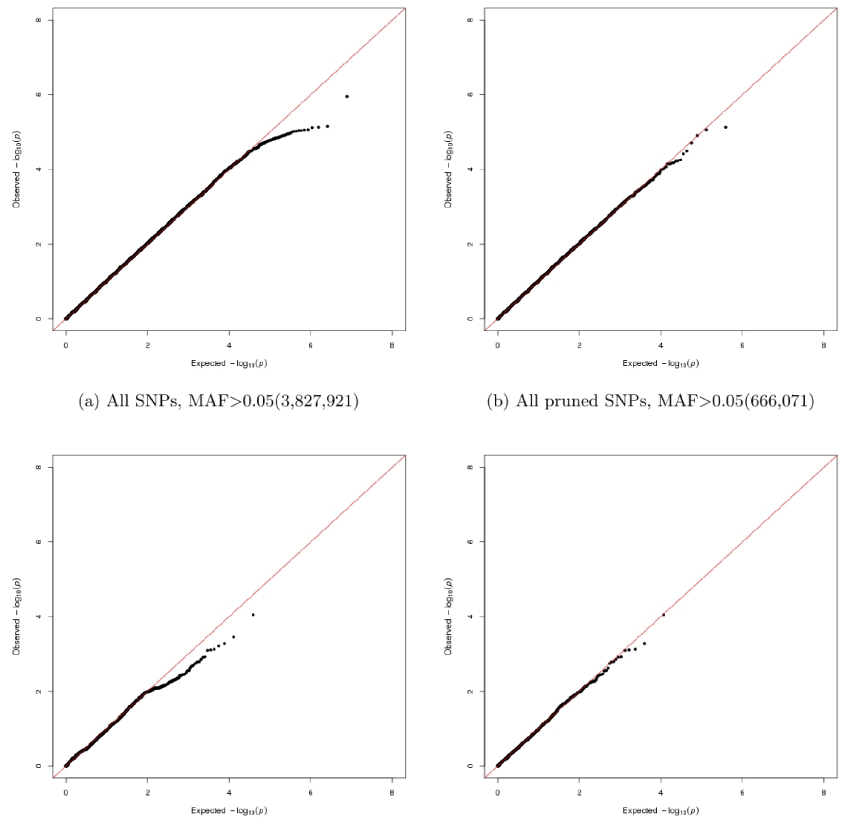
After quality control steps described in the Methods section, there were 5,852 brain eSNPs in SEED. Subsets of brain region specific eSNPs are also shown in Table 5.1. For example, there are 1,892 cerebellar cortex eSNPs in SEED, and in other brain regions, there are less than 1,000 pruned eSNPs. The number of all SNPs in SEED after LD pruning is 666,071.

After Bonferroni correction, none of the SNPs in any SNP subsets exceeded the corresponding adjusted significance level in SEED. However, among the temporal cortex eSNPs from the SEED GWAS, the top eSNP (rs9876540, $p = 8.8810^{-5}$) does reach statistical significance, while for the rest of SNP subsets in SEED, the p-values of top SNPs are about ten times greater than their adjusted significance level. Variant rs9876540 is located in the intron region of gene CDCP1 according to dbSNP,[139] and CDCP1 is transmembrane protein that involves in regulation of celluar events by RefSeq.[140]

5.3.2 Examining QQ-plots in SNP subsets of SEED GWAS

In SEED, we compared the distribution of SNP p-values between all SNPs and brain eSNPs, and before and after LD pruning in the QQ-plots in Figure 5.1. We can clearly see that before pruning, there is significant deflation in both all SNPs and brain eSNPs subset in SEED. After pruning, the deflation is greatly reduced, but no apparent positive signals can be observed above the x=y line in Figures 5.1b and 5.1d. This is consistent with results from Table 5.1 that the top SNPs are far from reaching statistical significance.

As for the brain region specific eSNP subsets in SEED, we also examined their QQ-plots in SEED in Figure 5.2. The QQ-plot on temporal cortex eSNP subset(Figure



(a) All SNPs, MAF>0.05(3,827,921)

(b) All pruned SNPs, MAF>0.05(666,071)

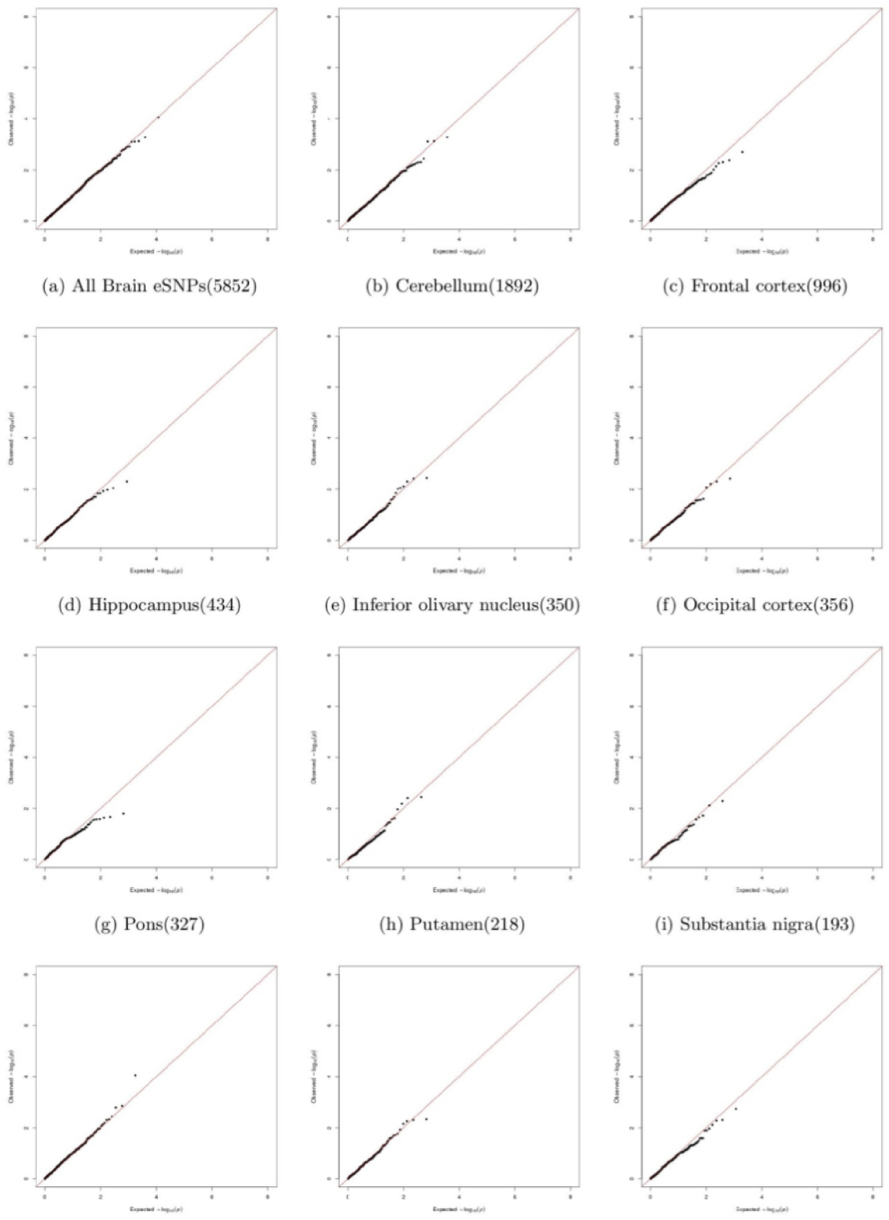
(c) Brain eSNPs, MAF>0.05(41,556)

(d) Brain pruned eSNPs, MAF>0.05(5,852)

Figure 1: SEED QQ plots comparing all SNPs vs. LD pruned SNPs(VIF), MAF>0.05

Figure 5.1: SEED QQ plots comparing all SNPs vs. LD pruned SNPs, MAF>0.05

5.2j)shows a potential positive signal above the x=y line without significant inflation or deflation, while the rest of QQ-plots from eSNP subsets show either deflation or no positive signal. This is also consistent with top SNP in temporal cortex eSNP subset from Table 5.1.



(a) All Brain eSNPs(5852) (b) Cerebellum(1892) (c) Frontal cortex(996)

(d) Hippocampus(434) (e) Inferior olivary nucleus(350) (f) Occipital cortex(356)

(g) Pons(327) (h) Putamen(218) (i) Substantia nigra(193)

(j) Temporal cortex(886) (k) Thalamus(343) (l) Intralobular white matter(574)

Figure 2: SEED QQ plots of pruned eSNPs from different brain tissues, MAF>0.05

Figure 5.2: SEED QQ plots comparing all SNPs vs. LD pruned SNPs, MAF>0.05

5.4 Discussion

We found that in SEED, the QQ-plot of temporal cortex eSNP subsets showed positive signal, and the top SNP rs9876540 in this subset almost reached the adjusted significance level in this subset. These results suggest that ASD may affect the gene expression in the temporal cortex or inferior olivary nucleus.

Our study reveals interesting findings in terms of possibly implicated brain regions in ASD. Brain imaging studies have showed that the temporal cortex, which plays a role in emotional regulation and social behavior,[47]might be affected in ASD patients.[141–144] However, there is currently no substantial evidence that the inferior olivary nucleus is involved in ASD, though some studies implicate this region in pediatric neurodegenerative disorders.[145] We need further studies that assess the association between structure or function of these brain regions and genetic profiles among ASD patients and controls. One of the major limitations in this study is that the brain eQTL studies often relied on a small sample of postmortem brains, and very few brain samples are from prenatal and postnatal subjects. This limits our coverage of brain eSNPs that may involve in brain development. The other limitation is that while LD pruning removes correlation between SNPs, possible significant signals may be lost in this process. In general, integrating brain eQTLs in GWAS studies, especially brain region eQTLs, can provide us some information on the etiology of psychiatric disorders. Future brain eQTLs studies with more samples from prenatal and postnatal subjects can help identify interesting brain eSNPs.

5.5 Supplementary material

Table 5.2: Source of Brain eQTLs

Author(year)	Sample description	Brain tissue source	Age Range	Brain region	Number of eSNPs	Genotyping Platform
Gibbs(2010)	150 neurologically normal caucasians	University of Maryland Brain Bank Johns Hopkins Hospital Baltimore Longitudinal Study of Aging	15-101	Cerebellum Frontal cortex Pons Temporal cortex	cis&trans: 8,888	Illumina Infinium HumanHap550 beadchip Illumina HumanRefseq-8 Expression BeadChip platform
Ramasamy(2014)	134 neurologically normal caucasians	Sun Health Research Institute Brain Bank Sudden Death Brain Bank in Edinburgh	16-102	Cerebellar cortex Frontal cortex Hippocampus Inferior olivary nucleus Occipital cortex Putamen Substantianigra Temporal cortex Thalamus Intralobular white matter	cis: 27,258	Affymetric Human Exon 1.0 ST array data Illumina genotyping platform
Myers(2007)	193 neurologically normal caucasians	National Institute of Aging Alzheimer Centers Miami Brain Bank	65-100	Pooled from 20% frontal cortex, 70% temporal cortex, 1% parietal cortex	cis: 443 trans: 336	Affymetric GeneChip Human Mapping 500K Array Set Illumina HumanRefseq-8 Expression BeadChip platform
Webster(2009)	AD caucasians: 176 non-AD caucasians: 188	National Institute of Aging Alzheimer Centers Miami Brain Bank	≥65	Pooled from frontal cortex (case:21%,control: 18%) temporal cortex (case:73%,control: 60%) parietal cortex (case:2%,control: 10%) cerebellar cortex (case:3%.control: 13%)	cis: 1,829 trans: 656	Affymetric GeneChip Human Mapping 500K Array Set Illumina HumanRefseq-8 Expression BeadChip platform
Zou(2012)	AD cerebellar: 197 AD temporal cortex: 202 non-AD cerebellar:	Mayo late-onset Alzheimer's disease(LOAD) GWAS	72 ± 6	Cerebellum Temporal cortex as validation	cis: 2,596	WG-DASL assays Illumina HumanHap300-Duo Genotyping BeadChips
Colantuoni(2011)	269 neurologically normal subjects 147 AA, 4 Asians, 112 Caucasians, 6 Hispanics	NIMH Brain Tissue Collection University of Maryland Brain & Tissue Bank	38 fetus, 18 infant(< 6 month), 15 child(1-10), 50 adolescents(10-20),148 adults (20-70)	Prefrontal Cortex	cis: 1,628	Illumina Infinium II 650K / Illumina Infinium HD Gemini 1M Duo BeadChips Illumina Beadchips

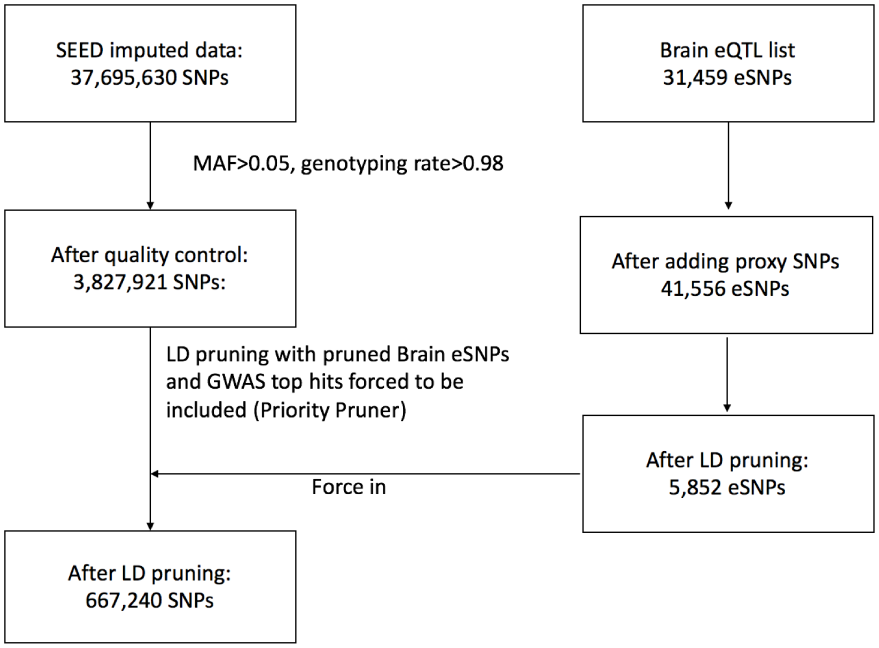


Figure 5.3: Flowchart

Chapter 6

Conclusions

6.1 Summary of conclusions

The epigenetics analyses on injection drug use revealed that there are some CpG sites that are suggestively associated with any injection drug use. The epigenetics analyses on specific types of injection drug use showed that heroin only injection and cocaine only injection lead to different epigenetics profiles changes. The joint analyses on four kinds of injection drug use identified groups of CpG sites that are shared across different injection drug use types, and these sites were enriched in immune response pathways.

Similarly, the HIV epigenetic analyses have confirmed a previously reported gene NLRC5 that was demethylated among chronic injection drug users. During the analyses, we used a new method to specifically controls for the cell composition problem

in HIV infection analyses. This analysis highlighted interesting genomic regions for downstream biological analyses on understanding chronic HIV infection.

The genetics analyses on ASD have shown potential ways on integrating expression data on brain into GWAS. With more data coming out on brain expression and larger GWAS, we will be able to conduct better functional enrichment analyses in GWAS and replicate our findings.

6.2 Future directions

With lowered cost on genotyping, sequencing and collection of epigenetic information, it is possible to collect this type of data in on-going epidemiologic cohorts and evaluate longitudinal changes on epigenetics data. It is both biologically interesting and public health relevant whether the epigenome changes over time and how much the changes accumulate at specific sites. We can evaluate both the epigenetic differences between quitting after long-term injection, and relapse after quitting. We treated these two injection status changes the same in our analyses, but they may lead to different epigenetic outcomes. With a balanced design with epigenetic data collected on injection, quitting, and relapse visits for every subject, we would be able to distinguish these two types of injection status changes.

Based on the epigenetics results, we can try to develop predictive models that utilize epigenetics information as biomarkers. The identified groups of CpG sites by the

joint analyses may also contribute to better feature selection for developing predictive models. We can use random forest to rank the importance of epigenetic features, and select the most predictive markers to develop machine learning prediction models by neural network, support vector machine, etc.

We would also test if there is overlap between methylation in the blood and brain tissue since the addiction process happens in the brain. We have proposed to collect epigenetics data on post-mortem brain tissue of opioid overdosed subjects to make comparison between top CpG sites in the brain and blood.

Since HIV infection is perfectly correlated with cell composition in the blood, alternatively, we can also do cell sorting first and then collect epigenetics data on specific types of blood cells to avoid cell composition problems. We can target CD4+ cell and CD8+ cell specifically since the chronic HIV infection mostly affects these two types of blood cells.

We can also improve our HIV analyses by measuring epigenetics data at the single cell level. Since the HIV virus is randomly integrated into the human genome, it might also affect the epigenome at the single cell level. We will be able to measure the variability across CD4+ cells and it may provide important target sites for HIV treatment.

Another important direction is that integrating genetics and expression data with epigenetics data, and to try to find interesting genomic regions informed by the integrated data. There are also other forms of functional genetics data, including histone

modification, DNA hypersensitivity sites and ChIP-seq. It is equally important to measure these types of data and integrate them with genetics and expression data.

Since the current brain eQTL database is established on about 100 brain tissues, we would need to expand the brain expression database to obtain more brain eQTLs. As long as the brain eQTLs and brain region specific eQTLs are well established, we will be able to get a more accurate picture on whether brain eQTLs are enriched in Autism Spectrum Disorder GWAS.

Bibliography

[1] S. W. G. of the Psychiatric Genomics Consortium *et al.*, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature*, vol. 511, no. 7510, pp. 421–427, 2014.

[2] B. L. Genberg, S. J. Gange, V. F. Go, D. D. Celentano, G. D. Kirk, and S. H. Mehta, "Trajectories of injection drug use over 20 years (1988–2008) in baltimore, maryland," *American journal of epidemiology*, vol. 173, no. 7, pp. 829–836, 2011.

[3] Z. Steel, C. Marnane, C. Iranpour, T. Chey, J. W. Jackson, V. Patel, and D. Silove, "The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013," *International journal of epidemiology*, vol. 43, no. 2, pp. 476–493, 2014.

[4] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters, "Lifetime prevalence and age-of-onset distributions of dsm-iv disorders

in the national comorbidity survey replication," *Archives of general psychiatry*, vol. 62, no. 6, pp. 593–602, 2005.

[5] W. W. Eaton, *Public mental health*. Oxford University Press, 2012.

[6] D. Vigo, G. Thornicroft, and R. Atun, "Estimating the true global burden of mental illness," *The Lancet Psychiatry*, vol. 3, no. 2, pp. 171–178, 2016.

[7] C. Wong, S. L. Odom, K. A. Hume, A. W. Cox, A. Fettig, S. Kucharczyk, M. E. Brock, J. B. Plavnick, V. P. Fleury, and T. R. Schultz, "Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review," *Journal of Autism and Developmental Disorders*, vol. 45, no. 7, pp. 1951–1966, 2015.

[8] W. S. Bush and J. H. Moore, "Genome-wide association studies," *PLoS computational biology*, vol. 8, no. 12, p. e1002822, 2012.

[9] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani *et al.*, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.

[10] R. M. Cantor, K. Lange, and J. S. Sinsheimer, "Prioritizing gwas results: a review of statistical methods and recommendations for their application," *The American Journal of Human Genetics*, vol. 86, no. 1, pp. 6–22, 2010.

[11] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five years of gwas discovery," *The American Journal of Human Genetics*, vol. 90, no. 1, pp. 7–24, 2012.

[12] C.-D. G. of the Psychiatric Genomics Consortium *et al.*, "Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis," *The Lancet*, vol. 381, no. 9875, pp. 1371–1379, 2013.

[13] P. Roussos, A. C. Mitchell, G. Voloudakis, J. F. Fullard, V. M. Pothula, J. Tsang, E. A. Stahl, A. Georgakopoulos, D. M. Ruderfer, A. Charney *et al.*, "A role for noncoding variation in schizophrenia," *Cell reports*, vol. 9, no. 4, pp. 1417–1429, 2014.

[14] A. Sekar, A. R. Bialas, H. de Rivera, A. Davis, T. R. Hammond, N. Kamitaki, K. Tooley, J. Presumey, M. Baum, V. Van Doren *et al.*, "Schizophrenia risk from complex variation of complement component 4," *Nature*, vol. 530, no. 7589, pp. 177–183, 2016.

[15] C. M. Bulik, P. F. Sullivan, and K. S. Kendler, "Genetic and environmental contributions to obesity and binge eating," *International Journal of Eating Disorders*, vol. 33, no. 3, pp. 293–298, 2003.

[16] P. Chaste and M. Leboyer, "Autism risk factors: genes, environment, and gene-environment interactions," *Dialogues in clinical neuroscience*, vol. 14, no. 3, p. 281, 2012.

[17] M. T. Tsuang, W. S. Stone, and S. V. Faraone, "Genes, environment and schizophrenia," *The British Journal of Psychiatry*, vol. 178, no. 40, pp. s18–s24, 2001.

[18] C. Ladd-Acosta and M. D. Fallin, "The role of epigenetics in genetic and environmental epidemiology," 2015.

[19] G. Egger, G. Liang, A. Aparicio, and P. A. Jones, "Epigenetics in human disease and prospects for epigenetic therapy," *Nature*, vol. 429, no. 6990, pp. 457–463, 2004.

[20] C. A. Cecil, E. Walton, and E. Viding, "Epigenetics of addiction: current knowledge, challenges, and future directions," *Journal of studies on alcohol and drugs*, vol. 77, no. 5, pp. 688–691, 2016.

[21] C. Ladd-Acosta, C. Shu, B. K. Lee, N. Gidaya, A. Singer, L. A. Schieve, D. E. Schendel, N. Jones, J. L. Daniels, G. C. Windham *et al.*, "Presence of an epigenetic signature of prenatal cigarette smoke exposure in childhood," *Environmental research*, vol. 144, pp. 139–148, 2016.

[22] Y. J. Loke, A. J. Hannan, and J. M. Craig, "The role of epigenetic change in autism spectrum disorders," *Frontiers in neurology*, vol. 6, 2015.

[23] S. Rangasamy, S. R. DMello, and V. Narayanan, "Epigenetics, autism spectrum,

and neurodevelopmental disorders," *Neurotherapeutics*, vol. 10, no. 4, pp. 742–756, 2013.

[24] B. M. Mathers, L. Degenhardt, B. Phillips, L. Wiessing, M. Hickman, S. A. Strathdee, A. Wodak, S. Panda, M. Tyndall, A. Toufik, and R. P. Mattick, "Global epidemiology of injecting drug use and hiv among people who inject drugs: a systematic review." *Lancet*, vol. 372, no. 9651, pp. 1733–1745, Nov 2008.

[25] K. A. Mack, "Illicit drug use, illicit drug use disorders, and drug overdose deaths in metropolitan and nonmetropolitan areasunited states," *MMWR. Surveillance Summaries*, vol. 66, 2017.

[26] C. for Behavioral Health Statistics, S. A. Quality, and M. H. S. Administration, "Key substance use and mental health indicators in the united states: Results from the 2016 national survey on drug use and health," in *HHS Publication No. SMA 17-5044, NSDUH Series H-52*. Substance Abuse and Mental Health Services Administration Rockville, MD, 2017.

[27] D. Vlahov, J. C. Anthony, A. Munoz, J. Margolick, K. E. Nelson, D. D. Celentano, L. Solomon, and B. F. Polk, "The alive study, a longitudinal study of hiv-1 infection in intravenous drug users: description of methods and characteristics of participants." *NIDA Res Monogr*, vol. 109, pp. 75–100, 1991.

[28] S. H. Mehta, G. D. Kirk, J. Astemborski, N. Galai, and D. D. Celentano, "Tem-

poral trends in highly active antiretroviral therapy initiation among injection drug users in baltimore, maryland, 1996–2008," *Clinical Infectious Diseases*, vol. 50, no. 12, pp. 1664–1671, 2010.

[29] A. Bargagli, A. Sperati, M. Davoli, F. Forastiere, and C. Perucci, "Mortality among problem drug users in rome: an 18-year follow-up study, 1980–97," *Addiction*, vol. 96, no. 10, pp. 1455–1463, 2001.

[30] S. M. Bird, S. J. Hutchinson, and D. J. Goldberg, "Drug-related deaths by region, sex, and age group per 100 injecting drug users in scotland, 2000–01," *The Lancet*, vol. 362, no. 9388, pp. 941–944, 2003.

[31] D. Vlahov, C.-l. Wang, N. Galai, J. Bareta, S. H. Mehta, S. A. Strathdee, and K. E. Nelson, "Mortality risk among new onset injection drug users," *Addiction*, vol. 99, no. 8, pp. 946–954, 2004.

[32] E. E. Schoenbaum, D. Hartel, P. A. Selwyn, R. S. Klein, K. Davenny, M. Rogers, C. Feiner, and G. Friedland, "Risk factors for human immunodeficiency virus infection in intravenous drug users," *New England Journal of Medicine*, vol. 321, no. 13, pp. 874–879, 1989.

[33] T. Kerr, B. D. L. Marshall, M.-J. Milloy, R. Zhang, S. Guillemi, J. S. G. Montaner, and E. Wood, "Patterns of heroin and cocaine injection and plasma hiv-1 rna suppression among a long-term cohort of injection drug users," *Drug*

and Alcohol Dependence, vol. 124, no. 1-2, pp. 108–112, 07 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3342432/

[34] K. E. Poundstone, R. E. Chaisson, and R. D. Moore, "Differences in hiv disease progression by injection drug use and by sex in the era of highly active antiretroviral therapy," *Aids*, vol. 15, no. 9, pp. 1115–1123, 2001.

[35] J. Kraus, "Expression and functions of μ-opioid receptors and cannabinoid receptors type 1 in t lymphocytes," *Annals of the New York Academy of Sciences*, vol. 1261, no. 1, pp. 1–6, 2012.

[36] P. Sacerdote, "Opioid-induced immunosuppression." *Curr Opin Support Palliat Care*, vol. 2, no. 1, pp. 14–18, Mar 2008.

[37] ——, "Opioids and the immune system," *Palliative medicine*, vol. 20, no. 8_suppl, pp. 9–15, 2006.

[38] S. C. McQuown and M. A. Wood, "Epigenetic regulation in substance use disorders," *Current psychiatry reports*, vol. 12, no. 2, pp. 145–153, 2010.

[39] D. A. Nielsen, A. Utrankar, J. A. Reyes, D. D. Simons, and T. R. Kosten, "Epigenetics of drug abuse: predisposition or response," *Pharmacogenomics*, vol. 13, no. 10, pp. 1149–1160, 2012.

[40] E. C. Prom-Wormley, J. Ebejer, D. M. Dick, and M. S. Bowers, "The genetic

epidemiology of substance use disorder: A review," *Drug and alcohol dependence*, 2017.

[41] G. Ebrahimi, G. Asadikaram, H. Akbari, M. H. Nematollahi, M. Abolhassani, G. Shahabinejad, L. Khodadadnejad, and M. Hashemi, "Elevated levels of dna methylation at the oprm1 promoter region in men with opioid use disorder," *The American journal of drug and alcohol abuse*, pp. 1–7, 2017.

[42] C. A. Cecil, E. Walton, and E. Viding, "Dna methylation, substance use and addiction: a systematic review of recent animal and human research from a developmental perspective," *Current Addiction Reports*, vol. 2, no. 4, pp. 331–346, 2015.

[43] X. Zhang, A. C. Justice, Y. Hu, Z. Wang, H. Zhao, G. Wang, E. O. Johnson, B. Emu, R. E. Sutton, J. H. Krystal *et al.*, "Epigenome-wide differential dna methylation between hiv-infected and uninfected individuals," *Epigenetics*, vol. 11, no. 10, pp. 750–760, 2016.

[44] K. N. Nelson, Q. Hui, D. Rimland, K. Xu, M. S. Freiberg, A. C. Justice, V. C. Marconi, and Y. V. Sun, "Identification of hiv infection-related dna methylation sites and advanced epigenetic aging in hiv-positive, treatment-naive us veterans," *Aids*, vol. 31, no. 4, pp. 571–575, 2017.

[45] G. P. Wang, A. Ciuffi, J. Leipzig, C. C. Berry, and F. D. Bushman, "Hiv integration site selection: analysis by massively parallel pyrosequencing reveals

association with epigenetic modifications," *Genome research*, vol. 17, no. 8, pp. 1186–1194, 2007.

[46] U. Mbonye and J. Karn, "Transcriptional control of hiv latency: cellular signaling pathways, epigenetics, happenstance and the hope for a cure," *Virology*, vol. 454, pp. 328–339, 2014.

[47] A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[48] D. L. Christensen, D. A. Bilder, W. Zahorodny, S. Pettygrove, M. S. Durkin, R. T. Fitzgerald, C. Rice, M. Kurzius-Spencer, J. Baio, and M. Yeargin-Allsopp, "Prevalence and characteristics of autism spectrum disorder among 4-year-old children in the autism and developmental disabilities monitoring network," *Journal of Developmental & Behavioral Pediatrics*, vol. 37, no. 1, pp. 1–8, 2016.

[49] M. Elsabbagh, G. Divan, Y.-J. Koh, Y. S. Kim, S. Kauchali, C. Marcín, C. Montiel-Nava, V. Patel, C. S. Paula, C. Wang *et al.*, "Global prevalence of autism and other pervasive developmental disorders," *Autism Research*, vol. 5, no. 3, pp. 160–179, 2012.

[50] T. A. Lavelle, M. C. Weinstein, J. P. Newhouse, K. Munir, K. A. Kuhlthau, and L. A. Prosser, "Economic burden of childhood autism spectrum disorders," *Pediatrics*, vol. 133, no. 3, pp. e520–e529, 2014.

[51] M. Stuart and J. H. McGrew, "Caregiver burden after receiving a diagnosis of an autism spectrum disorder," *Research in Autism Spectrum Disorders*, vol. 3, no. 1, pp. 86–97, 2009.

[52] J. P. Leigh and J. Du, "Brief report: Forecasting the economic burden of autism in 2015 and 2025 in the united states," *Journal of autism and developmental disorders*, vol. 45, no. 12, pp. 4135–4139, 2015.

[53] S. M. Myers, C. P. Johnson *et al.*, "Management of children with autism spectrum disorders," *Pediatrics*, vol. 120, no. 5, pp. 1162–1182, 2007.

[54] J. Hallmayer, S. Cleveland, A. Torres, J. Phillips, B. Cohen, T. Torigoe, J. Miller, A. Fedele, J. Collins, K. Smith *et al.*, "Genetic heritability and shared environmental factors among twin pairs with autism," *Archives of general psychiatry*, vol. 68, no. 11, pp. 1095–1102, 2011.

[55] S. De Rubeis and J. D. Buxbaum, "Genetics and genomics of autism spectrum disorder: embracing complexity," *Human molecular genetics*, vol. 24, no. R1, pp. R24–R31, 2015.

[56] S. Sandin, P. Lichtenstein, R. Kuja-Halkola, H. Larsson, C. M. Hultman, and A. Reichenberg, "The familial risk of autism," *Jama*, vol. 311, no. 17, pp. 1770–1777, 2014.

[57] K. Wang, H. Zhang, D. Ma, M. Bucan, J. T. Glessner, B. S. Abrahams,

D. Salyakina, M. Imielinski, J. P. Bradfield, P. M. Sleiman *et al.*, "Common genetic variants on 5p14. 1 associate with autism spectrum disorders," *Nature*, vol. 459, no. 7246, pp. 528–533, 2009.

[58] L. A. Weiss, D. E. Arking, M. J. Daly, A. Chakravarti, C. W. Brune, K. West, A. OConnor, G. Hilton, R. L. Tomlinson, A. B. West *et al.*, "A genome-wide linkage and association scan reveals novel loci for autism," *Nature*, vol. 461, no. 7265, pp. 802–808, 2009.

[59] R. Anney, L. Klei, D. Pinto, J. Almeida, E. Bacchelli, G. Baird, N. Bolshakova, S. Bölte, P. F. Bolton, T. Bourgeron *et al.*, "Individual common variants exert weak effects on the risk for autism spectrum disorders," *Human molecular genetics*, vol. 21, no. 21, pp. 4781–4792, 2012.

[60] A. Moreno-De-Luca, D. W. Evans, K. Boomer, E. Hanson, R. Bernier, R. P. Goin-Kochel, S. M. Myers, T. D. Challman, D. Moreno-De-Luca, M. M. Slane *et al.*, "The role of parental cognitive, behavioral, and motor profiles in clinical variability in individuals with chromosome 16p11. 2 deletions," *Jama psychiatry*, vol. 72, no. 2, pp. 119–126, 2015.

[61] E. B. Robinson, B. M. Neale, and S. E. Hyman, "Genetic research in autism spectrum disorders," *Current opinion in pediatrics*, vol. 27, no. 6, p. 685, 2015.

[62] C. A. Boyle, S. Boulet, L. A. Schieve, R. A. Cohen, S. J. Blumberg, M. Yeargin-Allsopp, S. Visser, and M. D. Kogan, "Trends in the prevalence of developmental

disabilities in us children, 1997–2008," *Pediatrics*, vol. 127, no. 6, pp. 1034–1042, 2011.

[63] J. L. Matson and M. Shoemaker, "Intellectual disability and its relationship to autism spectrum disorders," *Research in developmental disabilities*, vol. 30, no. 6, pp. 1107–1114, 2009.

[64] Y. Kim, K. Xia, R. Tao, P. Giusti-Rodriguez, V. Vladimirov, E. Van Den Oord, and P. Sullivan, "A meta-analysis of gene expression quantitative trait loci in brain," *Translational psychiatry*, vol. 4, no. 10, p. e459, 2014.

[65] A. Ramasamy, D. Trabzuni, S. Guelfi, V. Varghese, C. Smith, R. Walker, T. De, L. Coin, R. De Silva, M. R. Cookson *et al.*, "Genetic variability in the regulation of gene expression in ten regions of the human brain," *Nature neuroscience*, vol. 17, no. 10, pp. 1418–1428, 2014.

[66] D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, and N. J. Cox, "Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas," *PLoS genetics*, vol. 6, no. 4, p. e1000888, 2010.

[67] L. Chunyu, "Brain expression quantitative trait locus mapping informs genetic studies of psychiatric diseases."

[68] E. Courchesne, "Brain development in autism: early overgrowth followed by

premature arrest of growth," *Developmental Disabilities Research Reviews*, vol. 10, no. 2, pp. 106–111, 2004.

[69] H. C. Hazlett, M. Poe, G. Gerig, R. G. Smith, J. Provenzale, A. Ross, J. Gilmore, and J. Piven, "Magnetic resonance imaging and head circumference study of brain size in autism: birth through age 2 years," *Archives of general psychiatry*, vol. 62, no. 12, pp. 1366–1376, 2005.

[70] B. M. Mathers, L. Degenhardt, B. Phillips, L. Wiessing, M. Hickman, S. A. Strathdee, A. Wodak, S. Panda, M. Tyndall, A. Toufik *et al.*, "Global epidemiology of injecting drug use and hiv among people who inject drugs: a systematic review," *The Lancet*, vol. 372, no. 9651, pp. 1733–1745, 2008.

[71] G. M. Lucas, M. Griswold, K. A. Gebo, J. Keruly, R. E. Chaisson, and R. D. Moore, "Illicit drug use and hiv-1 disease progression: a longitudinal study in the era of highly active antiretroviral therapy," *American journal of epidemiology*, vol. 163, no. 5, pp. 412–420, 2006.

[72] S. A. Strathdee, T. B. Hallett, N. Bobrova, T. Rhodes, R. Booth, R. Abdool, and C. A. Hankins, "Hiv and risk environment for injecting drug users: the past, present, and future," *The Lancet*, vol. 376, no. 9737, pp. 268–284, 2010.

[73] S. J. Blumberg, M. D. Bramlett, M. D. Kogan, L. A. Schieve, J. R. Jones, and M. C. Lu, *Changes in prevalence of parent-reported autism spectrum disorder in school-aged US children: 2007 to 2011-2012*. US Department of Health and

Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2013, no. 65.

[74] C. Gillberg, M. Cederlund, K. Lamberg, and L. Zeijlon, "Brief report:the autism epidemic. the registered prevalence of autism in a swedish urban area," *Journal of autism and developmental disorders*, vol. 36, no. 3, p. 429, 2006.

[75] C. J. Newschaffer, M. D. Falb, and J. G. Gurney, "National autism prevalence trends from united states special education data," *Pediatrics*, vol. 115, no. 3, pp. e277–e282, 2005.

[76] P. F. Sullivan, "The psychiatric gwas consortium: big science comes to psychiatry," *Neuron*, vol. 68, no. 2, pp. 182–186, 2010.

[77] J. W. Ng, L. M. Barrett, A. Wong, D. Kuh, G. D. Smith, and C. L. Relton, "The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities," *Genome biology*, vol. 13, no. 6, p. 246, 2012.

[78] J. Madrigano, A. A. Baccarelli, M. A. Mittleman, D. Sparrow, P. S. Vokonas, L. Tarantini, and J. Schwartz, "Aging and epigenetics: longitudinal changes in gene-specific dna methylation," *Epigenetics*, vol. 7, no. 1, pp. 63–70, 2012.

[79] Y.-F. Chiu, A. E. Justice, and P. E. Melton, "Longitudinal analytical approaches to genetic data," *BMC genetics*, vol. 17, no. 2, p. S4, 2016.

[80] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware, *Applied longitudinal analysis*. John Wiley & Sons, 2012, vol. 998.

[81] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[82] G. E. Hoffman, J. G. Mezey, and E. E. Schadt, "lrgpr: interactive linear mixed model analysis of genome-wide association studies with composite hypothesis testing and regression diagnostics in r," *Bioinformatics*, vol. 30, no. 21, pp. 3134–3135, 2014.

[83] H. M. Kang, J. H. Sul, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti, E. Eskin *et al.*, "Variance component model to account for sample structure in genome-wide association studies," *Nature genetics*, vol. 42, no. 4, pp. 348–354, 2010.

[84] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman, "Fast linear mixed models for genome-wide association studies," *Nature methods*, vol. 8, no. 10, pp. 833–835, 2011.

[85] X. Zhou and M. Stephens, "Genome-wide efficient mixed-model analysis for association studies," *Nature genetics*, vol. 44, no. 7, pp. 821–824, 2012.

[86] S. L. Zeger, K.-Y. Liang, and P. S. Albert, "Models for longitudinal data: a generalized estimating equation approach," *Biometrics*, pp. 1049–1060, 1988.

[87] A. Agresti, *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015.

[88] U. Halekoh, S. Højsgaard, J. Yan *et al.*, "The r package geepack for generalized estimating equations," *Journal of Statistical Software*, vol. 15, no. 2, pp. 1–11, 2006.

[89] S. Moran, C. Arribas, and M. Esteller, "Validation of a dna methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences," *Epigenomics*, vol. 8, no. 3, pp. 389–399, 2016.

[90] C. S. Wilhelm-Benartzi, D. C. Koestler, M. R. Karagas, J. M. Flanagan, B. C. Christensen, K. T. Kelsey, C. J. Marsit, E. A. Houseman, and R. Brown, "Review of processing and analysis methods for dna methylation array data," *British journal of cancer*, vol. 109, no. 6, pp. 1394–1402, 2013.

[91] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry, "Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays," *Bioinformatics*, vol. 30, no. 10, pp. 1363–1369, 2014.

[92] J.-P. Fortin, T. J. Triche Jr, and K. D. Hansen, "Preprocessing, normaliza-

tion and integration of the illumina humanmethylationepic array with minfi," *Bioinformatics*, vol. 33, no. 4, pp. 558–560, 2016.

[93] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin, "Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis," *BMC bioinformatics*, vol. 11, no. 1, p. 587, 2010.

[94] D. L. McCartney, R. M. Walker, S. W. Morris, A. M. McIntosh, D. J. Porteous, and K. L. Evans, "Identification of polymorphic and off-target probe binding sites on the illumina infinium methylationepic beadchip," *Genomics data*, vol. 9, pp. 22–24, 2016.

[95] Y.-a. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, and R. Weksberg, "Discovery of cross-reactive probes and polymorphic cpgs in the illumina infinium humanmethylation450 microarray," *Epigenetics*, vol. 8, no. 2, pp. 203–209, 2013.

[96] E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey, "Dna methylation arrays as surrogate measures of cell mixture distribution," *BMC bioinformatics*, vol. 13, no. 1, p. 86, 2012.

[97] T. J. Triche Jr, D. J. Weisenberger, D. Van Den Berg, P. W. Laird, and K. D. Siegmund, "Low-level processing of illumina infinium dna methylation beadarrays," *Nucleic acids research*, vol. 41, no. 7, pp. e90–e90, 2013.

[98] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[99] A. E. Teschendorff, J. Zhuang, and M. Widschwendter, "Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies," *Bioinformatics*, vol. 27, no. 11, pp. 1496–1505, 2011.

[100] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, "The sva package for removing batch effects and other unwanted variation in high-throughput experiments," *Bioinformatics*, vol. 28, no. 6, pp. 882–883, 2012.

[101] J. A. Gagnon-Bartsch and T. P. Speed, "Using control genes to correct for unwanted variation in microarray data," *Biostatistics*, vol. 13, no. 3, pp. 539–552, 2012.

[102] J. Maksimovic, J. A. Gagnon-Bartsch, T. P. Speed, and A. Oshlack, "Removing unwanted variation in a differential methylation analysis of illumina humanmethylation450 array data," *Nucleic acids research*, vol. 43, no. 16, pp. e106–e106, 2015.

[103] A. E. Jaffe and R. A. Irizarry, "Accounting for cellular heterogeneity is critical in epigenome-wide association studies," *Genome biology*, vol. 15, no. 2, p. R31, 2014.

[104] E. A. Houseman, J. Molitor, and C. J. Marsit, "Reference-free cell mixture adjustments in analysis of dna methylation data," *Bioinformatics*, vol. 30, no. 10, pp. 1431–1439, 2014.

[105] A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry, "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies," *International journal of epidemiology*, vol. 41, no. 1, pp. 200–209, 2012.

[106] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[107] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.

[108] P. Geeleher, L. Hartnett, L. J. Egan, A. Golden, R. A. Raja Ali, and C. Seoighe, "Gene-set analysis is severely biased when applied to genome-wide methylation data," *Bioinformatics*, vol. 29, no. 15, pp. 1851–1857, 2013.

[109] B. Phipson, J. Maksimovic, and A. Oshlack, "missmethyl: an r package for analyzing data from illuminas humanmethylation450 platform," *Bioinformatics*, vol. 32, no. 2, pp. 286–288, 2015.

[110] S. Horvath, "Dna methylation age of human tissues and cell types," *Genome biology*, vol. 14, no. 10, p. 3156, 2013.

[111] K. Boulias, J. Lieberman, and E. L. Greer, "An epigenetic clock measures accelerated aging in treated hiv infection," *Molecular cell*, vol. 62, no. 2, pp. 153–155, 2016.

[112] S. Bhattacharjee, P. Rajaraman, K. B. Jacobs, W. A. Wheeler, B. S. Melin, P. Hartge, M. Yeager, C. C. Chung, S. J. Chanock, N. Chatterjee *et al.*, "A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits," *The American Journal of Human Genetics*, vol. 90, no. 5, pp. 821–835, 2012.

[113] P. F. OReilly, C. J. Hoggart, Y. Pomyen, F. C. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. Coin, "Multiphen: joint model of multiple phenotypes can increase discovery in gwas," *PloS one*, vol. 7, no. 5, p. e34861, 2012.

[114] Y. Wei, T. Tenzen, and H. Ji, "Joint analysis of differential gene expression in multiple studies using correlation motifs," *Biostatistics*, vol. 16, no. 1, pp. 31–46, 2014.

[115] N. Galai, M. Safaeian, D. Vlahov, A. Bolotin, and D. Celentano, "Longitudinal patterns of drug injection behavior in the alive study cohort, 1988–2000: description and determinants," *American Journal of Epidemiology*, vol. 158, no. 7, pp. 695–704, 2003.

[116] A. A. Lambert, G. D. Kirk, J. Astemborski, S. H. Mehta, R. A. Wise, and M. B. Drummond, "Hiv infection is associated with increased risk for acute exacerbation of copd," *Journal of acquired immune deficiency syndromes (1999)*, vol. 69, no. 1, p. 68, 2015.

[117] D. A. Nielsen, V. Yuferov, S. Hamon, C. Jackson, A. Ho, J. Ott, and M. J. Kreek, "Increased oprm1 dna methylation in lymphocytes of methadone-maintained former heroin addicts," *Neuropsychopharmacology*, vol. 34, no. 4, pp. 867–873, 2009.

[118] M.-R. Chao, D. Fragou, P. Zanos, C.-W. Hu, A. Bailey, S. Kouidou, and L. Kovatsi, "Epigenetically modified nucleotides in chronic heroin and cocaine treated mice," *Toxicology letters*, vol. 229, no. 3, pp. 451–457, 2014.

[119] L. G. Tsaprouni, T.-P. Yang, J. Bell, K. J. Dick, S. Kanoni, J. Nisbet, A. Viñuela, E. Grundberg, C. P. Nelson, E. Meduri *et al.*, "Cigarette smoking reduces dna methylation levels at multiple genomic loci but the effect is partially reversible upon cessation," *Epigenetics*, vol. 9, no. 10, pp. 1382–1396, 2014.

[120] B. R. Joubert, S. E. Håberg, R. M. Nilsen, X. Wang, S. E. Vollset, S. K. Murphy, Z. Huang, C. Hoyo, Ø. Midttun, L. A. Cupul-Uicab *et al.*, "450k epigenome-wide scan identifies differential dna methylation in newborns related to maternal

smoking during pregnancy," *Environmental health perspectives*, vol. 120, no. 10, p. 1425, 2012.

[121] C. Liu, R. Marioni, Å. K. Hedman, L. Pfeiffer, P. Tsai, L. Reynolds, A. Just, Q. Duan, C. Boer, T. Tanaka *et al.*, "A dna methylation biomarker of alcohol consumption," *Molecular psychiatry*, 2016.

[122] X. Zhang, Y. Hu, A. C. Justice, B. Li, Z. Wang, H. Zhao, J. H. Krystal, and K. Xu, "Dna methylation signatures of illicit drug injection and hepatitis c are associated with hiv frailty," *Nature communications*, vol. 8, no. 1, p. 2243, 2017.

[123] C. for Disease Control, Prevention *et al.*, "Hiv surveillance report, 2016," *Atlanta, GA: Centers for Disease Control and Prevention*, vol. 26, 2016.

[124] A. van Sighem, L. Gras, P. Reiss, K. Brinkman, F. de Wolf *et al.*, "Life expectancy of recently diagnosed asymptomatic hiv-infected patients approaches that of uninfected individuals," *Aids*, vol. 24, no. 10, pp. 1527–1535, 2010.

[125] S. G. Deeks, S. R. Lewin, and D. V. Havlir, "The end of aids: Hiv infection as a chronic disease," *The Lancet*, vol. 382, no. 9903, pp. 1525–1533, 2013.

[126] J. T. Maricato, M. N. Furtado, M. C. Takenaka, E. R. Nunes, P. Fincatti, F. M. Meliso, I. D. da Silva, M. G. Jasiulionis, M. C. de Araripe Sucupira, R. S. Diaz *et al.*, "Epigenetic modulations in activated cells early after hiv-1 infection and

their possible functional consequences," *PloS one*, vol. 10, no. 4, p. e0119234, 2015.

[127] U. Mbonye and J. Karn, "Control of hiv latency by epigenetic and non-epigenetic mechanisms," *Current HIV research*, vol. 9, no. 8, pp. 554–567, 2011.

[128] M. Verma, "Epigenetic regulation of hiv, aids, and aids-related malignancies," *Cancer Epigenetics: Risk Assessment, Diagnosis, Treatment, and Prognosis*, pp. 381–403, 2015.

[129] E. Ay, F. Banati, M. Mezei, A. Bakos, H. H. Niller, K. Buzás, and J. Minarovits, "Epigenetics of hiv infection: promising research areas and implications for therapy," *AIDS Rev*, vol. 15, no. 3, pp. 181–188, 2013.

[130] A. M. Gross, P. A. Jaeger, J. F. Kreisberg, K. Licon, K. L. Jepsen, M. Khosroheidari, B. M. Morsey, S. Swindells, H. Shen, C. T. Ng *et al.*, "Methylome-wide analysis of chronic hiv infection reveals five-year increase in biological age and epigenetic targeting of hla," *Molecular cell*, vol. 62, no. 2, pp. 157–168, 2016.

[131] R. P. Westergaard, T. Hess, J. Astemborski, S. H. Mehta, and G. D. Kirk, "Longitudinal changes in engagement in care and viral suppression for hiv-infected injection drug users," *AIDS (London, England)*, vol. 27, no. 16, p. 2559, 2013.

[132] A. Jaffe, "Flowsorted. blood. 450k: Illumina humanmethylation data on sorted blood cell populations," *R package version 1.16.0*, 2017.

[133] Y. Cheng, J. F. Quinn, and L. A. Weiss, "An eqtl mapping approach reveals that rare variants in the sema5a regulatory network impact autism risk," *Human molecular genetics*, vol. 22, no. 14, pp. 2960–2972, 2013.

[134] A. J. Myers, J. R. Gibbs, J. A. Webster, K. Rohrer, A. Zhao, L. Marlowe, M. Kaleem, D. Leung, L. Bryden, P. Nath *et al.*, "A survey of genetic human cortical gene expression," *Nature genetics*, vol. 39, no. 12, p. 1494, 2007.

[135] J. A. Webster, J. R. Gibbs, J. Clarke, M. Ray, W. Zhang, P. Holmans, K. Rohrer, A. Zhao, L. Marlowe, M. Kaleem *et al.*, "Genetic control of human brain transcript expression in alzheimer disease," *The American Journal of Human Genetics*, vol. 84, no. 4, pp. 445–458, 2009.

[136] J. R. Gibbs, M. P. van der Brug, D. G. Hernandez, B. J. Traynor, M. A. Nalls, S.-L. Lai, S. Arepalli, A. Dillman, I. P. Rafferty, J. Troncoso *et al.*, "Abundant quantitative trait loci exist for dna methylation and gene expression in human brain," *PLoS genetics*, vol. 6, no. 5, p. e1000952, 2010.

[137] C. Colantuoni, B. K. Lipska, T. Ye, T. M. Hyde, R. Tao, J. T. Leek, E. A. Colantuoni, A. G. Elkahloun, M. M. Herman, D. R. Weinberger *et al.*, "Temporal dynamics and genetic control of transcription in the human prefrontal cortex," *Nature*, vol. 478, no. 7370, p. 519, 2011.

[138] F. Zou, H. S. Chai, C. S. Younkin, M. Allen, J. Crook, V. S. Pankratz, M. M. Carrasquillo, C. N. Rowley, A. A. Nair, S. Middha *et al.*, "Brain expression genome-wide association study (egwas) identifies human disease-associated variants," *PLoS genetics*, vol. 8, no. 6, p. e1002707, 2012.

[139] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbsnp: the ncbi database of genetic variation," *Nucleic acids research*, vol. 29, no. 1, pp. 308–311, 2001.

[140] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D61–D65, 2006.

[141] R. T. Schultz, I. Gauthier, A. Klin, R. K. Fulbright, A. W. Anderson, F. Volkmar, P. Skudlarski, C. Lacadie, D. J. Cohen, and J. C. Gore, "Abnormal ventral temporal cortical activity during face discrimination among individuals with autism and asperger syndrome," *Archives of general Psychiatry*, vol. 57, no. 4, pp. 331–340, 2000.

[142] M. Zilbovicius, N. Boddaert, P. Belin, J.-B. Poline, P. Remy, J.-F. Mangin, L. Thivard, C. Barthélémy, and Y. Samson, "Temporal lobe dysfunction in childhood autism: a pet study," *American Journal of Psychiatry*, vol. 157, no. 12, pp. 1988–1993, 2000.

[143] K. Garbett, P. J. Ebert, A. Mitchell, C. Lintas, B. Manzi, K. Mirnics, and A. M.

Persico, "Immune transcriptome alterations in the temporal cortex of subjects with autism," *Neurobiology of disease*, vol. 30, no. 3, pp. 303–311, 2008.

[144] B. P. Ander, N. Barger, B. Stamova, F. R. Sharp, and C. M. Schumann, "Atypical mirna expression in temporal cortex associated with dysregulation of immune, cell cycle, and other pathways in autism spectrum disorders," *Molecular autism*, vol. 6, no. 1, p. 37, 2015.

[145] M. Mirabelli-Badenier, G. Morana, C. Bruno, M. Di Rocco, P. Striano, E. De Grandis, E. Veneselli, A. Rossi, and R. Biancheri, "Inferior olivary nucleus involvement in pediatric neurodegenerative disorders: does it play a role in neuroimaging pattern-recognition approach?" *Neuropediatrics*, vol. 46, no. 02, pp. 104–109, 2015.

Vita

Please see the next page.

Chang(April) Shu

Johns Hopkins Bloomberg School of Public Health
624 N. Broadway, Room 784, Baltimore, MD 21205
443-562-6397 aprilshu@jhu.edu

EDUCATION

Doctor of Philosophy (Ph.D.), Department of Mental Health **Expected 2018**
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Research area: Psychiatric and Behavioral Genetic Epidemiology
Advisors: Dr. Brion Maher & Dr. Margaret Daniele Fallin

Master of Health Science (M.H.S.), Department of Biostatistics **Expected 2018**
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Research area: Genomics
Advisor: Dr. Hongkai Ji

Master of Science (M.S.), Department of Epidemiology **May 2014**
Harvard School of Public Health, Boston, MA
Concentration: Neuro-psychiatric Epidemiology
Thesis: Examining the Influence of Substance Use Disorder Treatment on Smoking Cessation
Advisors: Dr. Benjamin Lê Cook & Dr. Edward Giovannucci

Bachelor of Science (B.S.), Department of Chemistry **July 2012**
Tsinghua University, Beijing, PR China
Senior Thesis: Growth characteristics of auxotrophic yeast and its application
Advisor: Dr. Yen Wei

RESEARCH EXPERIENCE

Graduate Research Assistant; Advisor: Dr. Brion Maher (Department of Mental Health) **2016-2017**
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Research Project: DNA methylation markers associated with injection drug use and HIV infection among chronic injection drug users in the ALIVE study (dissertation)

- Preprocessed and normalized DNA methylation data over 800 blood samples from Illumina HumanMethylationEPIC array with adjustment for batch effect.
- Performed DNA methylation single-site analyses to assess the effect of recent drug injection (cocaine, heroin and speedball injection) and HIV infection on changes in methylation levels across ~800,000 CpG sites by Generalized Estimating Equation(GEE) and linear mixed effect model.
- Generated DNA methylation age for each visit of the study subject and assessed the acceralation of aging between recent injection drug users and non-users, HIV positive and negative groups.
- Performed gene ontology analysis on genes associated with top ranked CpG sites.

Graduate Research Assistant; Advisor: Dr. Hongkai Ji (Department of Biostatistics) **2016-2017**
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Research Project: Joint analysis of DNA methylations in multiple types of injection drug use in the ALIVE study using correlation motifs (dissertation)

- Performed EM algorithm to extract correlation structure (referred as correlation motif) based on results from DNA methylation single-site analyses on four different phenotypes (any recent injection drug use, heroin injection, cocaine injection and speedball injection)
- Conducted permutation tests to obtain emprical test statistics distribution from single-site analyses.
- Obtained four meaningful correlation structure among four phenotypes and reranked the top CpG sites integrating information from the correlation motif.

Graduate Research Assistant; Advisor: Dr. Dani Fallin (Department of Mental Health) **2015-2017**
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Research Project: Integrating brain expression quantitative loci (eQTLs) in Autism Spectrum Disorder genome-wide association analyses

- Conducted Genome-wide Association analyses (GWAS) on ~1,200 Autism Spectrum Disorder (ASD) cases and population controls in the Study to Explore Early Development (SEED) study after linkage disequilibrium(LD) pruning in plink.
- Extracted ~40,000 brain eQTLs and brain region specific eQTLs from 6 eQTL studies.
- Integrated brain eQTLs SNPs with the ASD GWAS by subsetting GWAS SNPs to brain eQTL SNPs only and brain region specific eQTL SNPs only to increase the power to detect biologically meaningful SNPs associated with ASD. Examine and compare the QQ plots before and after subsetting.
- Applied similar analyses in Psychiatric Genomics Consortium Autism(PGC-AUT) panel.

Graduate Research Assistant; Advisor: Dr. Dani Fallin (Department of Mental Health) **2015-2016**
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Research Project: Prediction of prenatal cigarette smoke exposure by DNA methylation signature in blood samples from child

- Performed prediction of prenatal cigarette smoking by 26 CpG loci that were previously associated by machine learning methods (support vector machine, random forest, etc.).
- Data management and cleaning on the phenotype data in SAS.

Graduate Research Assistant; Advisor: Dr. Benjamin Cook (Harvard Medical School) **2012-2014**
Center for Multicultural Mental Health Research, Cambridge, MA
Research Project: Examining the association between substance use disorder treatment and smoking cessation (master's thesis)

- Extracted smoking cessation, substance use disorder treatment, and social demographic variables from 12,796 subjects in the National Survey on Drug Use and Health (2009-2011) by STATA.
- Assessed the association between use of substance use disorder treatment and smoking cessation by logistic regression.
- Imputed missing data for any covariates by multiple imputation.

PUBLICATIONS

Ladd-Acosta, **C., Shu**, C., Lee, B. K., Gidaya, N., Singer, A., Schieve, L. A., . . . Windham, G. C. (2016). Presence of an epigenetic signature of prenatal cigarette smoke exposure in childhood. *Environmental research, 144*, 139-148.

Shu, C., & Cook, B. L. (2015). Examining the association between substance use disorder treatment and smoking cessation. *Addiction, 110*(6), 1015-1024.

Cook, B. L., Wayne, G. F., Kafali, E. N., Liu, Z., **Shu, C.**, & Flores, M. (2014). Trends in smoking among adults with mental illness and association between mental health treatment and smoking cessation. *JAMA, 311*(2), 172-182.

Publication in progress:

Shu, C., Benke, K., Ladd-Acosta, C., Jaffe, A., Daniels, J.L., Newschaffer, C.J., Reynolds, A.M., Schendel, D.E., Schieve, L.A. , Fallin, M.D. Integrating Expression Quantitative Brain Loci in Autism Spectrum Disorder Genome-wide Association analyses.

Shu, C., Bakulski, K.M., Benke, K.S., Jaffe, A.E., Wang, S., Sabunciyan, S.H., Mehta, S.H., Kirk, G.D., Maher, B.S. DNA methylation markers associated with injection drug use status and HIV infection among chronic injection drug users in the ALIVE study

PRESENTATIONS

Conference Posters

Shu C., Maher B.S., Bakulski K.M., Benke K.S., Jaffe A.E., Wang S., Sabunciyan S., Mehta S., Kirk G. Circulating DNA methylation markers associated with injection status and HIV infection among chronic injection drug users in the ALIVE study. Presented at NIDA Genetics Consortium Meeting, December 2016, Bethesda, MD

Shu, C., Benke, K.S., Ladd-Acosta, C., Jaffe, A., Daniels, J.L., Newschaffer, C.J., Reynolds, A.M., Schendel, D.E., Schieve, L.A., Fallin, M.D. Integrating Expression Quantitative Brain Loci in Autism Spectrum Disorder Genome-wide Association analyses. Presented at the 2016 Annual Meeting of The American Society of Human Genetics, Octorber 2016, Vancouver, Canada.

Shu, C., Benke, K., Ladd-Acosta, C., Jaffe, A., Daniels, J.L., Newschaffer, C.J., Reynolds, A.M., Schendel, D.E., Schieve, L.A., Fallin, M.D. Integrating Expression Quantitative Brain Loci in Autism Spectrum Disorder Genome-wide Association analyses. Presented at 2016 International Society for Autism Research Meeting, May 2016, Baltimore, MD

Conference Talks

Shu, C., Bakulski, K.M., Benke, K.S., Jaffe, A.E., Wang, S., Sabunciyan, S.H., Mehta, S.H., Kirk, G.D., Maher, B.S. DNA methylation markers associated with injection drug use status and HIV infection among chronic injection drug users in the ALIVE study. Presented at NIH-CSSA Research Symposium, June 2017, Bethesda, MD

Shu, C., Benke, K.S., Ladd-Acosta, C., Jaffe, A., Daniels, J.L., Newschaffer, C.J., Reynolds, A.M., Schendel, D.E., Schieve, L.A., Fallin, M.D. Integrating Expression Quantitative Brain Loci in Autism Spectrum Disorder Genome-wide Association analyses. Presented at Research Potpourri of Department of Psychiatry and Behavioral Sciences, Johns Hopkins Hospital, May 2016, Baltimore, MD

TEACHING EXPERIENCE

Teaching Assistant, Introduction to Behavior and Psychiatric genetics, Dr. Peter Zandi **2017**
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health
- Taught a lecture on "Defining the Phenotype" during class for 50+ Masters and PhD students

Teaching Assistant, Data Analysis Workshops, Dr. Rick Thompson **2015**
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
- Held TA sessions and graded homeworks for a class of 50+ MPH students

HONORS & AWARDS

Reviewer's choice abstract in 2016 American Society of Human Genetics Meeting **2016**
- Abstract ranked top 10% of poster abstracts

Lucy Shum Memorial Scholarship for excellence in public health research, **2016**
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health
- Scholarship award for excellence in public health research

Delta Omega Scholarship Competition winners for Applied Research, **2015**
Johns Hopkins Bloomberg School of Public Health
- Research support for the project "Integration of Brain eQTLs and GWAS to identify genetic variants for Autism Spectrum Disorders" out of 90 applications.

School of Science Scholarship, Tsinghua University **2011**
- Awarded for excellence in academic performance in School of Science

Eternal Chemical Scholarship, Tsinghua University **2009**
- Awarded for excellence in academic performance in Department of Chemistry

SKILLS

- Computing: R, shell scripts, C/C++, SAS (Certified Clinical Trials Programmer), STATA, SPSS
- Language: English (bilingual proficiency); Mandarin (native)

VITA